STATISTICAL METHODS FOR HEALTH ACTUARIES
IBNR ESTIMATES: An Introduction

Jinadasa Gamage, Jed Linfield, Krzysztof Ostaszewski and Steven Siegel

October 31, 2007

Foreword

Over the years, I've heard from many health actuaries of their desire to incorporate more statistical concepts into their daily responsibilities, such as reserve estimates, benefit pricing, etc. At the same time, as a result of greater scrutiny on financial reports because of Sarbanes-Oxley and other measures, the pressure on health actuaries to demonstrate validity in their estimates has grown steadily.

Recognizing an opportunity to help serve its members in this age of increased financial oversight, the Health Section of the Society of Actuaries commissioned this series of guides on the use of statistical techniques specifically geared for the work of health actuaries. In this first guide in the series, the topic is an estimate well-known to health actuaries—the calculation of incurred but not reported (IBNR) health claims reserves. In particular, this guide focuses on the development of confidence intervals around IBNR estimates. Future guides to be published in this series include applications of credibility theory to health actuarial tasks and statistical approaches to prescription drug claim data.

The guides have been written with a number of distinct audiences in mind, and these audiences will likely want to use the guides differently. For this guide on IBNR calculations, an experienced health actuary with distant, yet pleasant (well, maybe not so pleasant) memories of actuarial exams may choose to skip over the introductory chapters and concentrate more on the later chapters. For beginning health actuaries, the statistical concepts in the guide may be fresh on their minds, but they might not yet have actually calculated an IBNR claim reserve. These actuaries can use the guide as an introduction to how IBNR claims reserves are typically calculated in practice and then move on to the statistical perspective.

Finally, for experienced health actuaries who have already incorporated statistical techniques into their daily practice, I hope this guide inspires them to further their work and devise new methods that they might want to share with the health actuarial community. Related to this, it should be noted that the techniques outlined in the guide represent a sample of those that can be used. The intent of the guide is neither to be all-inclusive of the variety of techniques applied in practice nor to demonstrate the absolute best or most advanced application of statistical theory for IBNR claims reserves calculations. Those topics, while worthy of research, are the subjects for other work and not the purpose of this guide.

The guide is divided into five sections along with an appendix with supplementary material. The following is a brief description of each section:

Section 1: Overview of Health Care Liabilities and Introduction to the Completion Factor Method
This section provides general information about health care liabilities and outlines a simplified example of how the completion factor method is commonly used in practice to calculate IBNR claims reserves. Experienced health actuaries will most likely find they will want to skip the material in this section.

Section 2: The Completion Factor Method Using Medical Insurance Data
In this section, the completion factor method is more fully explained with an expanded example. Typical refinements to the method are discussed, including the removal of outliers and other adjustments to the data. This section then begins to introduce a statistical perspective on the data and previews the technique to be used in Section 3.

Section 3: Regression Analysis for Recent Months and a Description of the Regression Workbook
This section provides an example of how regression can be used to calculate a confidence interval on the IBNR claim estimate for the two most recent months of incurred and paid claim data. Practical considerations of the claim data, such as benefit changes and different coverage levels, are discussed.

Section 4: Simulation Techniques to Estimate Confidence Intervals for IBNR Reserves
In this section, a step by step guide for employing simulation techniques to calculate confidence intervals and fitting probability distributions to the claim data is presented. For ease of application, the techniques in this section make use of an Excel add-in called @RISK. The techniques may be used with other simulation programs or Excel add-in software, but for illustrative purposes and to provide a completely workable example, detailed steps in @RISK are provided. As of the publication of this guide, a 10-day, free trial version of @RISK may be downloaded at www.palisade.com.

Section 5: Key Statistical Terms
For reference purposes, brief definitions of statistical terms and concepts used throughout the document are provided in this section.

Appendix
The appendix provides further reference material on multicolinearity to supplement the examples in the guide.

No introduction to this topic can be complete without mentioning the role of actuarial judgment. The techniques in the guide present the groundwork for producing IBNR estimates and confidence intervals. However, what cannot be taught in a guide is the experience that is gained over time through working with particular claim data and familiarity with an organization's business practices and unique circumstances. In this regard, any technique is only as good as the ultimate actuarial judgment that renders its results reasonable.

Acknowledgments

The Health Section Council would like to thank the authors commissioned to write this guide for their hard work in bringing this to completion: Jinadasa Gamage, Jed Linfield and Krzysztof Ostaszewski.

This guide would have not have been possible without the advice and wise counsel of the Project Oversight Group appointed for this effort: Rowen Bell, Elaine Canlas, David Dickson, Doug Fearrington, Chuck Fuhrer, Eric Smithback, Tony Wittman and Kurt Wrobel.

Special thanks to Walter James for contributing the techniques used in Section 4 and generously giving his time to review the material. Thanks also to Claire Bilodeau, Stuart Klugman and Jim Mange for their helpful comments on this material when it was presented as part of a Course 7 session.

Many actuarial students at Illinois State University at Normal contributed to the software and documentation. Significant contributions were made by the following students: Christine Doyle, Kathleen Hempe, Kritiga Muthuswamy, Maliha Niitsu and Christine Wiehe.

I feel fortunate to have participated as both a co-author and member of the Project Oversight Group for this guide. It is my hope that this guide sparks continued interest in this topic and that when health actuaries think Monte Carlo, it's for more than casinos and famous celebrity sightings.

Steven Siegel
SOA Research Actuary
October 2007

Introduction

In this document, we illustrate how to use statistical methods to estimate unknown values and create confidence intervals in the context of setting reserves for health (other than disability) insurance claims that have been incurred but not reported (IBNR).

Two general statistical approaches used in developing estimates and corresponding confidence intervals are:

- Statistical methods. For instance, a regression analysis might be applied to estimate a range for the value of claims liability.

- Additional knowledge of factors outside of the statistical method. For instance, the range calculated statistically might be adjusted to reflect a known change in benefit plan design.

It is possible to use both approaches on the same problem. However, it is important to have a framework for combining both approaches.

This guide also includes discussion of the most common non-statistical method used in the health care industry—the completion factor method.

This document is organized into the following sections:

- Section 1: Overview of Health Care Liabilities and Introduction to the Completion Factor Method
- Section 2: The Completion Factor Method Using Medical Insurance Data
- Section 3: Regression Analysis for Recent Months and a Description of the Regression Workbook
- Section 4: Simulation Techniques to Estimate Confidence Intervals for IBNR Reserves
- Section 5: Key Statistical Terms
- Appendix I: Multicolinearity

A note on terminology usage: For the sake of clarity and familiarity to readers, the terms *confidence interval* or simply *interval* are primarily used to describe our calculation of interest throughout the text. From a statistical point of view, the term *prediction interval* is more accurate to describe this calculation in most circumstances in this guide and in particular, regression. For interested readers, Section 5 contains the technical definitions of *confidence intervals* and *prediction intervals*.

Section 1—Overview of Health Care Liabilities and Introduction to the Completion Factor Method

In this section, we give an overview of health care liabilities and a description of one of the well-known methods for computing the health plan liability: the completion factor method.

1.1 Overview of Health Care Liabilities

1.1.1 Definitions

We will refer to the insured under a health insurance contract under consideration as the *member* of the health insurance plan. Members seeking medical care incur a cost that may be reimbursable by the insurance company (a *claim*). The month in which a member sees a provider for medical care is called the *incurred month*.

The amount ultimately paid for claims incurred in a given month is modeled by a process called *development*. One common method of modeling the development process is called the *chain ladder* method. This method is applied by estimating the ratios of amounts paid in consecutive months (*development factors*) or percentages of ultimate cost paid up to a given date (*completion factors*). These factors are related to one another in a way that will be illustrated in the examples that follow.

After a claim is incurred, it is submitted to the insurance company that is providing health care coverage for the claimant. The month during which the claim is reported to the insurance company is called the *reported* month, and the month in which the insurance company pays the claim is called the *paid month*. In this document, we are only concerned with the incurred and paid months.

The paid month comes after the incurred month, and it can, of course, occur no earlier than the incurred month. (In this document, we are not considering the issue of pre-paid medical care, or capitation.) In a report of paid claims by incurred month, each incurred month is given a row and each paid month is given a column (or each incurred month is given a column and each paid month is given a row). Only cells in which the paid month is equal to or later than the incurred month are nonzero. Since the nonzero values form a triangle, this report is typically called a *triangle report.*

Lag is the measure of the difference between incurred month and paid month. For instance, a claim that is incurred in July 2005 and paid in July 2005 is defined as being paid at lag 0. A claim that is incurred in July 2005 and paid in August 2005 has a lag of 1.

The process of calculating insurance liabilities is referred to as *valuation.* Unless stated otherwise, when we are using a month as a valuation date, we are referring to the last day of the month, which is the date as of which valuations are usually performed.

1.1.2 Types of Health Reserves

Health (other than disability) insurance reserves may be classified broadly into the following categories:

- *Policy reserves*: amounts necessary for contract obligations created by future claims.

- *Expense liabilities*: amounts needed to pay loss adjustment expenses, taxes and other expenses related to liabilities incurred by the insurer from operations prior to the statement date.

- *Claims reserves*: amounts needed to pay claims already incurred but not yet reported or paid.

In this document, we will be concerned with claims reserves only.

U.S. statutory reserving practices for health insurance in general are governed by the National Association of Insurance Commissioners' (*NAIC) Model Minimum Reserve Standards for Individual and Group Health Insurance.*

1.1.3 Claims Reserves

Claims reserves represent estimates of the amounts that the insurer expects to pay in the future on claims that have been incurred prior to the end of the reporting period. Three major types of claims reserves are: claims due and unpaid, claims in the course of settlement and claims that have been incurred but not reported. Claims due and unpaid are usually small and known with reasonable precision. The two other types of claims reserves, either because the amount due to the provider of medical care (or policyholder) has not been determined, or because the insurance company has no knowledge of the claims yet, produce a significant level of uncertainty requiring actuarial analysis. The classic methodology for estimating health claims reserves is similar to the methodology used in property/casualty insurance loss reserving. There are, however, notable differences in the data. Key differences include the following:

- Claim frequency. In property/casualty insurance, the policyholder rarely makes a claim. In many types of health insurance, there is a high probability that an individual policyholder will make a claim. Consequently, the number of claims that a health insurer must process and manage is staggering compared to the number processed and managed by a property/casualty insurer.

- Claim characteristics. Due to the enormous volume of claims, specific details, helpful in determining the expected amount of individual claims, are usually not known by the health valuation actuary. The exception to this may be large catastrophic claims, such as claims for severe burns and for low-weight premature infants.

- Contractual information. An actuary sometimes has knowledge of hospital contract details, which would help derive expected payment rates for specific known claims.

- Payment patterns. Health insurance claims develop much more quickly than many types of casualty claims, such as liability and workers' compensation. The development of health claims from unknown to known (to the insurer) and paid is measured in months. For some casualty products, the development from unknown to known takes years.

1.2 Introduction to the Completion Factor Method

Suppose we know the following about claims incurred in August 2005:

Exhibit 1.1
Claims Incurred, August 2005

Paid Month	Lag	Amount Paid
August 2005	0	$2,000
September 2005	1	1,000
October 2005	2	1,000
November 2005	3	400
December 2005	4	1,100
Total Incurred		$5,500

In addition, we know that all claims incurred in August 2005 have been paid by the end of December 2005. Based on this information, we can derive development factors and completion factors for the month of August 2005 as follows:

Exhibit 1.2
Derivation of Development and Completion Factors
August 2005

(1) Lag	(2) Amount Paid	(3) Cumulative Paid	(4) Development Factor	(5) Derivation of (4)	(6) Completion Factor	(7) Derivation of (6)
0	$2,000	$2,000			36.4% (1)	(3) Lag 0/(3) Lag 4
1	1,000	3,000	1.50 (2)	(3) Lag 1 ÷ (3) Lag 0	54.5%	(3) Lag 1/(3) Lag 4
2	1,000	4,000	1.33	(3) Lag 2 ÷ (3) Lag 1	72.7%	(3) Lag 2/(3) Lag 4
3	400	4,400	1.10	(3) Lag 3 ÷ (3) Lag 2	80.0%	(3) Lag 3/(3) Lag 4
4	1,100	5,500	1.25	(3) Lag 4 ÷ (3) Lag 3	100.0%	Assumed

(1) 36.4% = $2,000 / $5,500
(2) 1.50 = $3,000 / $2,000

Exhibit 1.2 shows the relationship between development factors and completion factors. Given an assumption about when claims are complete and a set of development factors, the corresponding completion factors can be calculated.

Now suppose we have the following information about claims incurred between August and December 2005 and paid through the end of December 2005.

Exhibit 1.3a
Claims Incurred August – December 2005, Paid Through December 2005

Incurred Month	Paid Month					Total
	Aug-05	Sep-05	Oct-05	Nov-05	Dec-05	
Aug-05	$2,000	$1,000	$1,000	$400	$1,100	$5,500
Sep-05		$2,000	1,800	1,400	800	$6,000
Oct-05			$3,000	3,000	2,000	$8,000
Nov-05				900	600	$1,500
Dec-05					$5,000	$5,000

Rather than use the row to indicate the incurred month and the column to indicate the paid month, it is possible to do the opposite, as in Exhibit 1.3b. The particular choice between the two formats is somewhat arbitrary, but one format may be more convenient depending on the further calculations to be performed.

Exhibit 1.3b
Claims Incurred August – December 2005, Paid Through December 2005

Paid Month	Incurred Month				
	Aug-05	Sep-05	Oct-05	Nov-05	Dec-05
Aug-05	$2,000				
Sep-05	1,000	$2,000			
Oct-05	1,000	1,800	$3,000		
Nov-05	400	1,400	3,000	$ 900	
Dec-05	1,100	800	2,000	600	$5,000
Total	$5,500	$6,000	$8,000	$1,500	$5,000

Exhibit 1.4 below shows how we can apply the completion factors from Exhibit 1.2 to the claims paid to date in Exhibit 1.3a in order to estimate claims reserves for IBNR as of the end of December 2005.

Exhibit 1.4
IBNR Claims Reserves
As of December 2005

(1) Incurred Month	(2) Lag Through Dec-05	(3) Claims Paid Through Dec-05	(4) Completion Factor	(5) = (3) ÷ (4) Estimated Incurred	(6) = (5) – (3) IBNR
Aug-05	4	$ 5,500	100.0%	$ 5,500	$ -
Sep-05	3	6,000	80.0%	7,500	1,500
Oct-05	2	8,000	72.7%	11,000	3,000
Nov-05	1	1,500	54.5%	2,750	1,250
Dec-05	0	5,000	36.4%	13,750	8,750
Total		$26,000		$40,500	$14,500

The completion factors used to calculate IBNR in Exhibit 1.4 are based solely on claims incurred in August 2005. It is clear, however, that every month does not develop in the same way. For example, the lag 1 development factor based on September 2005 data is 1.90, not 1.50 as we derived for August 2005.

As an additional example, if we were to calculate the development factors using data from September 2005 only, we would obtain the following development factors, based on the data from Exhibit 1.3b:

Lag 1 Development Factor = ($2,000 + $1,800) ÷ $2,000 = 1.90
Lag 2 Development Factor = ($2,000 + $1,800 + $1,400) ÷ $3,800 = 1.37
Lag 3 Development Factor = $6,000 ÷ $5,200 = 1.15

Moreover, there is a great deal of information about the development process implicit in the development observed through December 2005 of claims incurred during September, October November and December 2005. The chain ladder method enables us to take advantage of this additional information.

Exhibit 1.5 rearranges and calculates the cumulative claims by incurred month and lag from the information found in Exhibit 1.3b above.

Exhibit 1.5
Cumulative Paid Claims by Incurred Month and Lag
August – December 2005

Lag	Incurred Month				
	Aug-05	Sep-05	Oct-05	Nov-05	Dec-05
0	$2,000	$2,000	$3,000	$ 900	$5,000
1	3,000	3,800	6,000	1,500	
2	4,000	5,200	8,000		
3	4,400	6,000			
4	5,500				

Exhibit 1.6 below shows an intermediate step in the chain ladder method. It calculates cumulative sums across different combinations of lags and incurred months.

Exhibit 1.6
Chain Ladder Method
Intermediate Cumulative Sums

Lag	Incurred Months			
	Aug-05	Aug-05 - Sep-05	Aug-05 - Oct-05	Aug-05 - Nov-05
0				$ 7,900
1			$12,800	$14,300
2		$ 9,200 (1)	$17,200	
3	$4,400	$10,400 (2)		
4	$5,500			

(1) $9,200 = $4,000 + $5,200
(2) $10,400 = $4,400 + $6,000

Exhibit 1.7 uses the intermediate sums found in Exhibit 1.6 to calculate estimates of the development factors based upon the development observed through December 2005 of claims incurred between August and December 2005.

Exhibit 1.7
Development Factor Estimates
Claims Incurred August – December 2005, Paid Through December 2005

Lag	Development Factors Based on Incurred Months				Derivation
	Aug-05	Aug-05 - Sep-05	Aug-05 - Oct-05	Aug-05 - Nov-05	
1				1.81	Lag 1 ÷ Lag 0
2			1.34		Lag 2 ÷ Lag 1
3		1.13			Lag 3 ÷ Lag 2
4	1.25				Lag 4 ÷ Lag 3

Finally, Exhibit 1.8 translates the estimated development factors from Exhibit 1.7 into estimated completion factors using the method shown in Exhibit 1.2 above.

Exhibit 1.8
Estimated Completion Factors

(1) Lag	(2) Development Factor	(3) Completion Factor	(4) Derivation
0		29.1%	(3) Lag 1 ÷ (2) Lag 1
1	1.81	52.7%	(3) Lag 2 ÷ (2) Lag 2
2	1.34	70.8% (1)	(3) Lag 3 ÷ (2) Lag 3
3	1.13	80.0% (2)	(3) Lag 4 ÷ (2) Lag 4
4	1.25	100.0%	Assumed

(1) 70.8% = 80.0% divided by 1.13
(2) 80.0% = 100.0% divided by 1.25

Based on the completion factors we just derived and our calculation method in Exhibit 1.4, we estimate claims reserves (IBNR) as follows:

Exhibit 1.9
Estimated IBNR

Incurred Month	Lag Through Dec-05	Claims Paid Through Dec-05	Completion Factor	Estimated Incurred	IBNR
Aug-05	4	$ 5,500	100.0%	$ 5,500	$ -
Sep-05	3	6,000	80.0%	7,500	1,500
Oct-05	2	8,000	70.8%	11,304	3,304
Nov-05	1	1,500	52.7%	2,848	1,348
Dec-05	0	5,000	29.1%	17,185	12,185
Total		$26,000		$44,338	$18,338

For most business and regulatory purposes, an actuary would like to ultimately estimate an IBNR that is sufficient with high probability. A sufficient IBNR is a value for the liability calculated such that, once all of the claims are paid, the IBNR was greater than or equal to the actual outstanding claims paid. The sufficiency of the estimate that is ultimately recorded for accounting purposes can be measured using a confidence interval.

In succeeding sections, we further explore the calculation of a confidence interval for an IBNR estimate, starting with the completion factor method and then using other techniques.

Section 2—The Completion Factor Method Using Medical Insurance Data

In Section 1, we illustrated a simple example of the completion factor method. In this section, we expand that example to a 36-month data set.

This section is organized as follows:

1. We describe the data set for the 36-month example.
2. We discuss the determination of outliers.
3. We derive completion factors using the completion factor method.
4. We analyze the completion factors and describe limitations of the completion factor method.

2.1 The Data Set

Our data set is expanded from the example in Section 1 in the following ways:

- Instead of five months of data, we use 36 months of data.
- In addition to paid claims, we have membership data by incurred month.

Our data set consists of claims for medical coverage for the time period January 2001 through December 2003. The claims data:

- Is summarized in a paid vs. incurred month grid (triangle).
- Represents medical coverage with no deductibles and coinsurance and relatively low co-payments such as $10 per office visit. Thus we assume that the same percentage of claims is paid by the insurer in each month.
- Excludes prescription drug claims. Prescription drug claims usually have a shorter lag pattern than other medical claims.

The paid total claims data set can be found in the worksheet entitled "OriginalData" in the file called "Statistical Methods – Regression Method – Workbook.xls", available on the Web page of the SOA Web site containing this document.

The observation in each cell of the paid by incurred month grid is the sum of many individual transactions. Each transaction can typically be described as one of the following:

- Non-catastrophic initial claims—claims of smaller amounts when they are first paid by the claims department. Such claims are often auto-adjudicated (that is, no human intervention), so there is relatively little lag between the report date and the paid date.
- Catastrophic initial claims—larger amount claims that have on average a longer payment lag because of adjudication complexities and may cause significant changes in total IBNR.
- Adjustments—adjustments to existing reported claims for any new information or administrative errors.

With sufficient data, each of these transaction types could be modeled separately. As a simplification, we will treat the transactions as though they come from a single distribution.

The lag time for when initial claims are paid is dependent on:

- how quickly providers (e.g., hospitals, physicians) or claimants submit claims following incurral; and
- the speed at which the insurer processes and pays claims.

An adjustment occurs when a claim is paid and later the insurance company determines that a different amount is needed. Two of the key reasons for adjustments are:

- Claims paid in error—for instance, a claim was paid twice.
- Third party liability—these occur typically through coordination of benefits, subrogation and other similar contractual provisions.

When negative adjustments exceed the combination of initial claims and positive adjustments, an incurred/paid month cell will show a negative amount.

As the processes underlying initial claims and subsequent adjustments differ, their underlying statistical distributions are different; thus, by combining these processes, we add additional variation to our eventual model. This is particularly true at the longer lags, such as lags 9, 10 and 11 in our sample data.

The variation in adjustments is more important at later durations because:

- adjustments cannot be made until after initial claims are paid;

- initial claims payments become significantly less frequent at later durations.

For each month, we have corresponding membership information. We truncated the data set to assume that all claims are paid within 12 months (e.g., all claims incurred in January 2001 are paid by January 31, 2002). We truncated the data set because the vast majority of medical claims are paid within the first 12 months following incurral; thus, the applicability of our models to real life IBNR calculations is not significantly impacted by this truncation.

2.2 Outliers

An outlier is a claim amount that is outside the normal or expected range of values. The determination of an outlier depends on the model or method used to estimate IBNR. In this regard, an outlier with respect to one model may not necessarily be an outlier relative to another model. A general two-step procedure for determining and handling outliers is as follows:

First, graph the data to visually inspect which points may be considered as candidates for outliers. In many cases, a visual inspection will be enough to determine any claims that are obvious outliers. If a visual inspection does not reveal outliers, another method is to apply a statistical test such as the six-sigma rule. This rule states that if a point is within three standard deviations of the mean, it is not considered an outlier. This test can be applied easily by using the Standard Deviation function in Excel or another similar program.

Second, either remove the outlier completely from the data or adjust it to a value that would be more within the expected range of claim values. It is also helpful to calculate the IBNR with and without the outlier to determine its ultimate impact on the calculation and how significant it is. For instance, we may decide for the purposes of the calculation that an outlier that impacts the calculation by 5 percent or more is significant and requires action, but less than that requires no adjustment. If the outlier is removed completely, for most purposes, it is advisable to make a final adjustment to the calculated IBNR to account for the claim amount removed.

It is important when identifying outlier claims to clearly document the justification for why a claim has been identified as such and how the outlier was treated in the IBNR calculation method. This can become particularly relevant if revisions or adjustments are needed to the IBNR estimate that is eventually recorded for accounting purposes.

Adjustment for outliers and catastrophic cases is further discussed in Sections 3.5.1 and 3.6.

2.2.1 Screening Our Data for Outliers

In analyzing our data for potential outliers, we plot our data as in Exhibit 2.1. In that graph, t is the lag and i is the incurred month. "Original Y" is the claims amount incurred in month i, $0 \leq i \leq 35$, and paid at lag t, $0 \leq t \leq 12$. This plot shows that there is a potential outlier in the middle of the data set. We have identified this as the observation corresponding to $i = 12$ and $t = 7$. This point corresponds to the $756,000 that was incurred in January 2002 and paid in August 2002.

The graph in Exhibit 2.1 helps identify positive outliers, whereas Exhibit 2.2 illustrates all of the negative values in the data.

Exhibit 2.1
Initial Scatter Plot of the Data

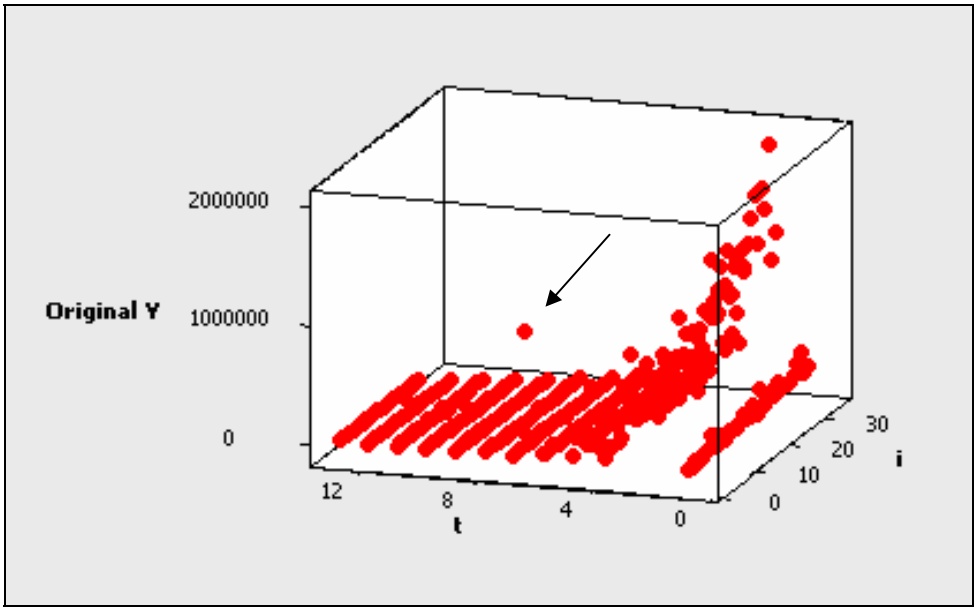
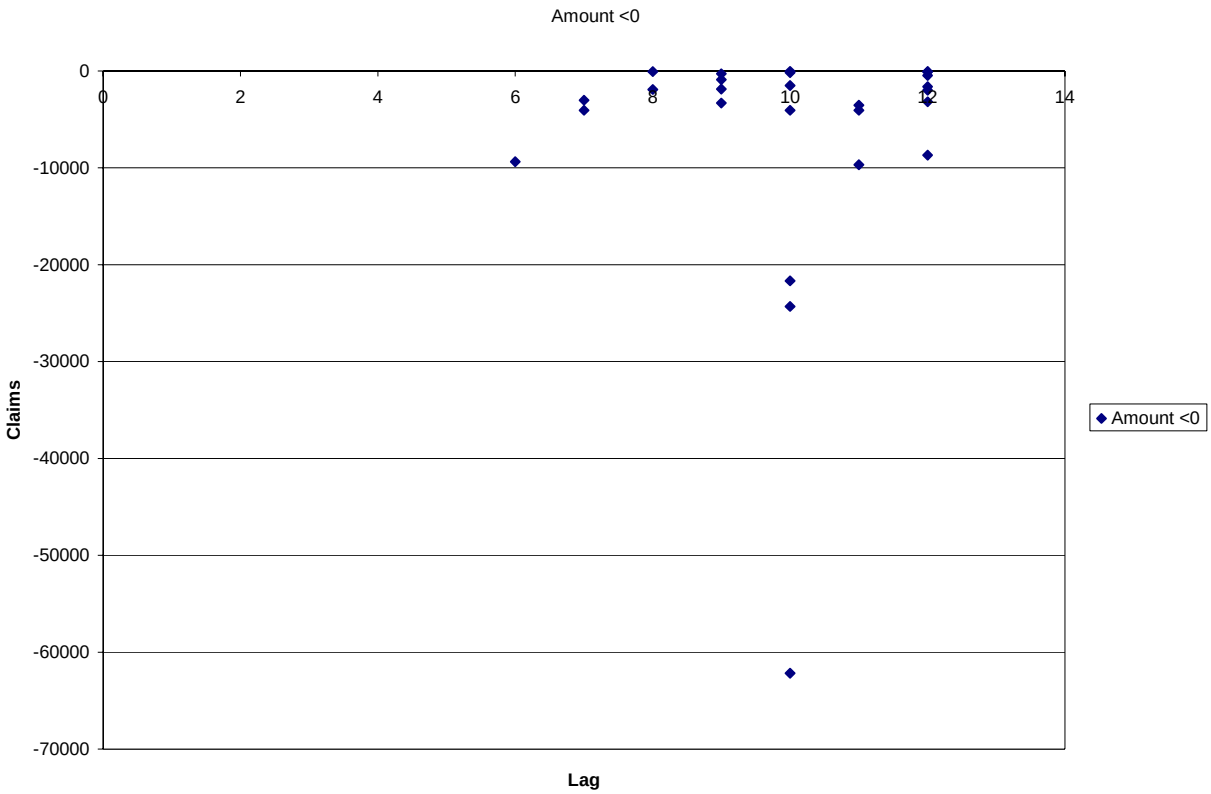


Exhibit 2.2
Data Points with Negative Values



The graph in Exhibit 2.2 shows negative data points by lag. The graph shows no negative data points in lags 0 through 5. We see a potential outlier at lag 10. The actual value is -$62,165, which occurs at i = 19 (month). This point is the sum of all claims transactions that were incurred in August 2002 and paid in June 2003. For this example, our next step would be to further investigate the outliers and consult with the claims processing department, as needed.

A consultation with the claims processing department reveals the following information:

- There was a $750,000 severe burn case that was incurred in January 2002 and paid in August 2002.
- The health insurance carrier paid, in September 2002, $60,000 for an auto accident claim that was incurred in August 2002. An auto insurance company reimbursed the health insurance carrier $60,000 for this claim in June 2003.

Based on this information, if we remove these claims as outliers, the new values in the triangle would be the following:

Exhibit 2.3
Potential Outliers

			(1)	(2)	(3) = (1) – (2)	
Incurred Month	Paid Month	Lag	Current Value in Claims Triangle	Potential Outlier Claim	New Value in Claims Triangle	Cause of Claim
Jan-02	Aug-02	7	$756,000	$750,000	$6,000	Severe Burn Case
Aug-02	Jun-03	10	($62,165)	($60,000)	($2,165)	Auto Accident—Third party payment received after claim paid

When we use the completion factor method, these are the two data points we will assess to determine their significance as outliers and how they will be treated for the calculation.

2.3 The Completion Factor Method

In Exhibit 2.4, we calculate the completion factors as we did in Exhibit 1.8 for the example in Section 1. The calculation is as follows:

Exhibit 2.4
Calculation of Completion Factors
36-month Sample Data

Completion Factor – Lag12	1.00000	(1) Due to data truncation as discussed above
Lag 12 claims (January 2001 through December 2002 – **24 months**)	$43,464,941 (*)	(2)
Lag 11 claims (January 2001 through December 2002 – **24 months**)	$43,428,968	(3)
Completion Factor – Lag 11	0.99917	(4) = (3) ÷ (2)
Lag 11 claims (January 2001 through January 2003 – **25 months**)	$45,977,288	(5)
Lag 10 claims (January 2001 through January 2003 – **25 months**)	$45,861,250	(6)
Lag 10 claims divided by Lag 11 Claims	0.99748	(7) = (6) ÷ (5)
Completion Factor – Lag 10	0.99665	(8) = (4) × (7)

(*) The workbook (named "Statistical Methods – Regression Method – Workbook.xls") has IBNR calculated with a $750,000 outlier excluded. The value $42,714,941, which is $750,000 less than $43,464,941, can be obtained by summing cells P6 through P29 in Tab WorkData2.

In a manner similar to Exhibit 1.4 in Section 1, we calculate IBNR by lag as follows:

Exhibit 2.5
Calculation of IBNR—Sample Data
Lags 0 Through 12
No Outliers Removed

Incurred Month	Lag	(1) Cumulative Paid Through December '03	(2) Completion Factor	(3)=(1)÷(2) Estimated Incurred Claims	(4)=(3)-(1) IBNR
Dec-03	0	$ 96,378	0.03936	$2,448,572	$2,352,193
Nov-03	1	1,283,817	0.59821	2,146,086	862,269
Oct-03	2	2,193,388	0.84701	2,589,571	396,184
Sep-03	3	2,688,921	0.91445	2,940,472	251,551
Aug-03	4	2,466,086	0.94775	2,602,055	135,969
Jul-03	5	2,502,042	0.96126	2,602,876	100,834
Jun-03	6	2,196,919	0.97130	2,261,842	64,923
May-03	7	2,385,024	0.99008	2,408,921	23,897
Apr-03	8	2,237,437	0.99275	2,253,778	16,341
Mar-03	9	2,361,919	0.99583	2,371,810	9,891
Feb-03	10	2,187,349	0.99665	2,194,699	7,351
Jan-03	11	2,548,319	0.99917	2,550,430	2,111
Dec-02	12	1,860,925	1.00000	1,860,925	0
Total—Lags 0 Through 12		$27,008,524		$31,232,037	$4,223,513
Total—Lags 2 Through 12		25,628,328		26,637,379	1,009,051

All of the incurred estimates for lags 2 through 11 are between $2 million and $3 million. The incurred claim estimates for lags 0 and 1, $2,448,572 and $2,146,085, respectively, seem to be reasonable estimates at first glance. However, as we will describe in later sections, the variation in these estimates is relatively large.

Returning to the evaluation of outliers, we previously identified two potential outliers:

- the $750,000 claim incurred in January 2002 and paid in August 2002;
- the ($60,000) claim adjustment incurred in August 2002 and paid in June 2003.

In Exhibit 2.6, we illustrate the impact of these potential outliers on IBNR for lags 2 through 12.

Exhibit 2.6
Effect of Potential Outliers on IBNR for Lags 2 through 12
Completion Factor Method

	With Both Outliers Included	With $750,000 Claim Excluded	With ($60,000) Claim Adjustment Excluded
IBNR	$1,009,051	$833,796	$1,033,963
Change in IBNR Due to Outlier	N/A	(17.4%)	2.5%

For this document, we assume that an absolute change, i.e., positive or negative, of 5 percent or more in IBNR is significant.

Based on that criterion, in analyzing the IBNR for lags 2 through 12, we conclude that the $750,000 claim is an outlier and the ($60,000) claim adjustment is not an outlier. In the regression workbook accompanying this guide, we have removed the $750,000 claim. In practice, there are several ways to treat this claim for IBNR purposes. Among possible approaches would be to include it in the experience during the factor development process, but to adjust it to a pre-determined amount consistent with an internal pooling arrangement, external reinsurance agreement or other reasonably expected amount. For the purposes of illustrating the methods in this guide, we have chosen to remove it completely from the experience during the factor development process.

When excluding the $750,000 burn case, our IBNR by month is as displayed in Exhibit 2.7.

Exhibit 2.7
IBNR with Catastrophic Burn Case Excluded

		(1)	(2)	(3) = (1) ÷ (2)	(4) = (3) – (1)
		Cumulative		Estimated	
Incurred		Paid through	Completion	Incurred	
Month	Lag	December '03	Factor	Claims	IBNR
Dec-03	0	$ 96,378	0.03990	$2,415,373	$2,318,994
Nov-03	1	1,283,817	0.60644	2,116,988	833,171
Oct-03	2	2,193,388	0.85865	2,554,461	361,073
Sep-03	3	2,688,921	0.92702	2,900,603	211,682
Aug-03	4	2,466,086	0.96077	2,566,775	100,689
Jul-03	5	2,502,042	0.97447	2,567,585	65,543
Jun-03	6	2,196,919	0.98465	2,231,175	34,255
May-03	7	2,385,024	0.98993	2,409,296	24,272
Apr-03	8	2,237,437	0.99263	2,254,040	16,603
Mar-03	9	2,361,919	0.99576	2,371,975	10,056
Feb-03	10	2,187,349	0.99659	2,194,823	7,475
Jan-03	11	2,548,319	0.99916	2,550,467	2,148
Dec-02	12	1,860,925	1.00000	1,860,925	-
		IBNR			
		Dec-02 – Oct-03 (Lags 2 through 12)			$ 833,796
		Nov-03 – Dec-03 (Lags 0 and 1)			3,152,165
		Total IBNR Dec-02 – Dec-03			$3,985,962

2.3.1 Analysis of the Completion Factor Method

As can be observed from Exhibit 2.7, over 79 percent of the IBNR ($3,152,165/$3,985,962) is represented by lags 0 and 1. Given this observation and the relative concentration of the IBNR, it is useful to analyze further the variation in the cumulative completion factors that were developed and in particular review the estimates for the most recent lags.

In this section, we will

- Demonstrate an approach for using descriptive statistics to analyze the variation in the cumulative completion factors to help assess the reasonability of the factors.
- Conclude that the completion factor method does not produce consistently reliable estimates for the most recent lag months based on our test of acceptable variability.

To test the cumulative completion factors, we analyze the unweighted monthly completion factors derived for each of the completed months, January 2001 through December 2002. Thus we have 24 data points on which to analyze the 13 completion factors developed (for lags 0 through 12, inclusive).

When comparing our cumulative completion factors with the mean of the unweighted monthly completion factors (after eliminating the $750,000 outlier, as described before), we obtain the following comparison:

Exhibit 2.8
Comparison of Completion Factors

Lag	Cumulative Factors	Mean of Unweighted Monthly Factors
0	0.03990	0.03215
1	0.60644	0.58789
2	0.85865	0.84632
3	0.92702	0.92366
4	0.96077	0.95917
5	0.97447	0.97211
6	0.98465	0.98437
7	0.98993	0.98975
8	0.99263	0.99223
9	0.99576	0.99551
10	0.99659	0.99645
11	0.99916	0.99904
12	1.00000	1.00000

The derivation of the mean of unweighted individual months is shown in a calculation in Tab "Unweighted Completion Factors" in the workbook "Statistical Methods – Regression Method – Workbook.xls." Because the values for each lag in Exhibit 2.8 are relatively close, we conclude that we can use the factor data from individual months to gain additional insight into the cumulative completion factors derived from using weighted averages.

Note that it might not always be the case that the monthly unweighted completion factors provide useful observations about the weighted average (cumulative) completion factors. For example, if the values in columns in Exhibit 2.8 appear to have large differences or show erratic patterns, then further analysis may not prove meaningful. For this approach, actuarial judgment will play a role in the ultimate value of the information.

In analyzing the monthly unweighted completion factors, a good measure of their stability is the standard deviation of the monthly unweighted completion factors divided by their mean for each lag. This ratio is analogous to a statistical measure known as the coefficient of variation.

Using the completion factor data from our 24 completed months, we obtain the results in Exhibit 2.9.

Exhibit 2.9
Completion Factor Method
Calculation of the Coefficient of Variation

	Completion Factor		
Lag	Mean of Individual Months	Standard Deviation	Coefficient of variation (Std. Dev./Mean)
0	0.03215	0.02548	0.79237
1	0.58789	0.12032	0.20467
2	0.84632	0.06114	0.07225
3	0.92366	0.04414	0.04778
4	0.95917	0.02483	0.02588
5	0.97211	0.02271	0.02336
6	0.98437	0.01471	0.01494
7	0.98975	0.01410	0.01425
8	0.99223	0.01444	0.01455
9	0.99551	0.01460	0.01467
10	0.99645	0.01025	0.01028
11	0.99904	0.00293	0.00294
12	1.00000	-	-

Note that the coefficient of variation at lag 12 is 0 since we assumed that all claims are paid by lag 12. In other words, we assumed that no claims are paid after lag 12 with 100 percent probability and no variation. Therefore, the standard deviation of the amount paid after lag 12 is 0, resulting in the same value for the coefficient of variation.

For the purposes of this document, we define a completion factor with acceptable variability as one with a coefficient of variation less than 0.1. This definition is based on our objective for stability of the completion factor estimate. Depending on the purpose of the analysis, a different threshold for the coefficient of variation might be used. Exhibit 2.9 indicates that, based on our definition, the completion factor method is producing estimates with unacceptable variability for lags 0 and 1. If we had selected 0.05 as our threshold, we would have concluded that the completion factor method is producing estimates with unacceptable variation for lags 0, 1 and 2.

An alternative illustration of the variation in the cumulative completion factors is to pose the following question: For every $100 of claims already paid, how much is the corresponding IBNR?

Exhibit 2.10 provides an example of the calculation implied by this question if the completion factor were 0.75.

Exhibit 2.10
Example of How Claims Paid and
Completion Factor Determine IBNR

Claims Paid	$100	(1)
Completion Factor	0.75	(2)
Incurred Claims	$133.33	(3) = (1) ÷ (2)
IBNR = Incurred Claims Minus Claims Paid	33.33	(4) = (3) –(1)

Applying this calculation to our set of completion factors yields the values shown in Exhibit 2.11.

Exhibit 2.11
IBNR per $100 of Paid Claims
for the Calculated Completion Factors

Lag	(1) Completion Factor*	(2) = $100÷(1) - $100 IBNR for $100 in Paid Claims	(3) Coefficient of Variation
0	0.03990	$2,406.14	0.79237
1	0.60644	64.90	0.20467
2	0.85865	16.46	0.07225
3	0.92702	7.87	0.04778
4	0.96077	4.08	0.02588
5	0.97447	2.62	0.02336
6	0.98465	1.56	0.01494
7	0.98993	1.02	0.01425
8	0.99263	0.74	0.01455
9	0.99576	0.43	0.01467
10	0.99659	0.34	0.01028
11	0.99916	0.08	0.00294
12	1.00000	-	-

* With catastrophic case excluded

As can be seen in comparing Exhibits 2.10 and 2.11, the coefficient of variation decreases as the "IBNR for $100 in Paid Claims" decreases (except in lags 8 and 9, where there is a slight increase). Again, we conclude that the variation is concentrated in the most recent lag months and the estimates for the most recent lag months have unacceptable variability.

2.3.2 Limitations of the Completion Factor Method

Limitations of the completion factor method include the following:

- The method, as presented, is cumulative; thus if incurred month *a* has twice as many claims as incurred month *b,* then the claims payment pattern of month *a* will have twice the weight in determining the completion factors as the claims payment pattern of month *b*. (Note that, in practice, alternatives to a cumulative approach may be applied.)
- For a given value of incurred claims, as more claims are paid, IBNR should decrease. In the completion factor method, incurred claims are determined by claims paid to date divided by completion factors. If the speed at which claims are paid increases, IBNR will increase even when incurred claims have not increased.
- The method uses a chain ladder approach to model the claim development process. Because no underlying probability distribution is defined, the method does not easily lend itself to the derivation of statistical measures such as confidence intervals.

2.3.3 Summary and Objective of Subsequent Sections

Now that we have completed an expanded example of the completion factor method including the treatment of outliers, it is helpful to review our ultimate goal in the context of what we know about the method. First, the objective of this guide is to describe techniques for calculating confidence intervals for IBNR estimates. In order to do so, we need to define an underlying probability distribution for either the claims data or the factors that are calculated by the completion factor method.

The completion factor method is considered a deterministic, as opposed to stochastic, method because, as it is applied, it results in only one estimated value for a particular completion factor, rather than a distribution of values. Because of this characteristic of the method, it does not easily lend itself to comprehensively defining an underlying probability distribution. Another characteristic of the completion factor method, as was illustrated in Section 2.3.1, is that much of the variation in the IBNR estimate is typically concentrated in the most recent lag months. The problem at hand is thus to use the known characteristics of the completion factor method in combination with other techniques to develop confidence intervals.

The remainder of this guide approaches this problem from a couple of perspectives. Section 3 presents an approach whereby regression is used to estimate the IBNR for the most recent lag months combined with a standard application of the completion factor method for the remaining prior months. Using regression allows us to define confidence intervals for the estimates on the most recent lag months.

Section 4 introduces simulation techniques to allow us to calculate a confidence interval for the entire IBNR estimate using the standard completion factor method as well as other approaches.

Section 3—Regression Analysis for Recent Months and a Description of the Regression Workbook

As described in Section 2, the completion factor method is considered a deterministic approach as it yields single value estimates, as opposed to a distribution of values. In this section, we combine regression techniques with the standard completion factor method. Furthermore, as demonstrated in Section 2, much of the estimate variation of the completion factor method is concentrated in the most recent lag months. For this reason, regression will be the approach for the IBNR estimates for lags 0 and 1 with the standard completion factor used for all other lags.

This section is organized as follows:

1. Description of the problem and variables.
2. Discussion of key statistical concepts used.
3. Description of the models we have in the regression workbook which we use in analyzing our data.
4. Detailed instructions on how to use the regression workbook.
5. Derivation of incurred claim costs using the regression workbook.
6. Explanation of how to calculate IBNR.
7. Confidence intervals
8. Expanding our analysis to nonuniform claim costs
9. Summary

3.1 Our Problem and Variables

To put our problem in the context of a regression analysis, the monthly claims data can be considered in terms of data points used to derive a regression line. Each point can be described by the month that claims were incurred and the estimate of total incurred claims for that month. Exhibit 3.1 presents a table of these data points. An example data point in the table would be: Lag Month 10, Incurred Claim Estimate $2,194,823. The incurred claim estimates for Lag Months 2-12 were previously derived in Section 2 by applying the completion factor method. The remaining incurred claim estimates come directly from the data set because we assume all claims are paid within 12 months after they were incurred. Exhibit 3.1 does not show data for Lag Months 0 and 1 because we will be using regression later in this section to estimate those values.

Exhibit 3.1 also provides the per member per month (PMPM) incurred claim estimate for each lag month. The PMPM is simply calculated by dividing the total incurred claim estimate for a particular month by the number of members covered that month. The PMPM incurred claim estimate is commonly used in practice as a way to normalize the level of claim activity that occurs from changes in plan membership.

With this perspective in mind, there are 34 data points—one each for months 0 through 33 (January 2001 through October 2003)—in our data set. Further extending this regression perspective, the primary independent variable for each point is time (the month incurred) and the dependent variable is the total or PMPM incurred claim estimate. Since time is the primary independent variable, Column 3 in Exhibit 3.1, "Months in Regression," specifies the chronological sequence of the months in the regression model. For instance, January 2001, the earliest month in our data set, is labeled Month 0 in Column 3.

Exhibit 3.1
PMPM Incurred Claim Estimates from Completion Factor Method

(1)	(2)	(3)	(4)	(5)	(6)=(4)÷(5)
Lag	Incurred Month	Months in Regression	Estimated Incurred Claims	Membership	PMPM Estimate
2	Oct-03	33	$2,554,461	11,843	$215.69
3	Sep-03	32	2,900,603	11,731	247.26
4	Aug-03	31	2,566,775	11,689	219.59
5	Jul-03	30	2,567,585	11,787	217.83
6	Jun-03	29	2,231,175	11,814	188.86
7	May-03	28	2,409,296	11,927	202.00
8	Apr-03	27	2,254,040	11,986	188.06
9	Mar-03	26	2,371,975	12,130	195.55
10	Feb-03	25	2,194,823	12,201	179.89
11	Jan-03	24	2,550,467	12,227	208.59
12	Dec-02	23	1,860,925	12,132	153.39
13	Nov-02	22	1,762,655	11,951	147.49
14	Oct-02	21	2,034,275	11,889	171.11
15	Sep-02	20	1,699,016	11,735	144.78
16	Aug-02	19	1,859,121	11,655	159.51
17	Jul-02	18	2,103,032	11,577	181.66
18	Jun-02	17	1,872,651	11,580	161.71
19	May-02	16	1,933,155	11,703	165.18
20	Apr-02	15	1,974,315	11,654	169.41
21	Mar-02	14	1,589,754	11,753	135.26
22	Feb-02	13	1,715,552	11,823	145.10
23	Jan-02	12	1,843,543	11,705	157.50
24	Dec-01	11	1,410,154	11,555	122.04
25	Nov-01	10	1,673,063	11,444	146.20
26	Oct-01	9	1,852,522	11,456	161.71
27	Sep-01	8	1,587,443	11,400	139.25
28	Aug-01	7	1,915,574	11,420	167.74
29	Jul-01	6	1,722,416	11,180	154.06
30	Jun-01	5	1,755,594	11,174	157.11
31	May-01	4	1,602,252	11,130	143.96
32	Apr-01	3	1,610,332	11,069	145.48
33	Mar-01	2	1,962,246	11,070	177.26
34	Feb-01	1	1,765,964	11,118	158.84
35	Jan-01	0	1,609,389	11,154	144.29

The data from Exhibit 3.1 can be found in Tab "Workdata3" in our workbook, "Statistical Methods – Regression Method – Workbook.xls."

There are two other independent variables which will be analyzed:

- Weekday/weekend indicator (Day Factor). Many medical offices are closed (or have limited hours) on weekends and holidays while hospitals and other facilities remain open for emergency and urgent care. We assume that on an average weekend/holiday, 35 percent of the utilization of an average weekday is experienced. The effect of including this variable in the model is to normalize the number of weekdays per month.

- Unspecified time variable (Time 2). There is the capacity in the regression workbook to add this additional time variable for each month. This variable has been purposely left undefined. It may be assigned values of 0 or 1 and would be incorporated into the regression model if its inclusion improves the model from a statistical standpoint. As well, its inclusion in the model should be reasonable from a practical understanding of the data. For instance, if we knew that all health care providers increased their rates by 10 percent on January 1, 2002, we could set this variable equal to 0 for months in 2001 and 1 for months in 2002 and 2003. This variable is different from the primary independent variable in the sense that it is assigned one of only two possible values, whereas the primary variable can take on a myriad of values depending on the unit of time we use to measure time.

3.2 Key Statistical Concepts

Before applying regression directly to IBNR estimation, the following key statistical concepts relating to regression are reviewed in this section for the benefit of the reader:

- R-square and adjusted R-square values;
- Number of data points;
- p-values in a regression output;
- Interval estimation;
- Interval estimates for simple linear regression;
- Residuals.

3.2.1 R-Square and Adjusted R-Square Values

When regression models are run, among the first results reviewed are the R-square and adjusted R-square values.

R-square is the coefficient of determination of the regression model. This measure is interpreted as the percentage of the variation in the observed values of the dependent variable that is explained by the regression model. The larger the value of R-square, the greater is the indication that the model is satisfactory. R-square is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

with

$$SS_T = \sum_i (y_i - \bar{y})^2, SS_R = \sum_i (\hat{y}_i - \bar{y})^2, SS_E = \sum_i (y_i - \hat{y}_i)^2.$$

In the above, *SSR* is the sum of squares accounted for in the regression model, *SST* is the total sum of squares of the deviations of the dependent variable values from their mean, and *SSE = SST - SSR*.

One drawback of R-square is that adding more independent variables to the model causes the R-square value to increase even though the improvement to the model, if any, may not justify the addition. In order to correct for this, the adjusted R-square is defined to be

$$R^2(adj) = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)},$$

where *k* is the number of independent variables (also known as regressors) in the regression model and *n* is the total number of observations on the dependent variable. Larger adjusted R-square values indicate a better fit of the model.

3.2.2 Number of Data Points

The number of data points can impact the R-square value. In our example, there are 34 data points. Let's say that the best model we can derive with 34 points has an adjusted R-square value of *x*. Furthermore, let's assume that if we use 30 data points, we can derive an adjusted R-square value of *y*, with *y* being 1 percent greater than *x*. Should we use the model with 30 data points? If we take this reduction in the number of points to the extreme, R-square and adjusted R-square would equal their maximum value of 1 if we had only two data points.

We follow the principle that, unless we have a compelling reason to exclude points, such as with outliers, we do not eliminate data points just to produce a model with a slightly better fit. However, we often transform variables or add additional variables to produce a better fit.

3.2.3 p-Values in a Regression Output

For a regression analysis, there are several associated test statistics and corresponding p-values. These are normally included as output from a statistical package. The F-statistic and corresponding p-value test if the overall model is statistically significant. A p-value less than 0.05 is an indicator that the overall model is statistically significant when compared to a model using none of the independent variables. The t-statistics and corresponding p-values test each individual regression coefficient (beta value). The independent (predictor) variables with p-values less than 0.05 are considered to be useful in describing the relationship between the independent (predictor) variables and the dependent (response) variable.

3.2.4 Interval Estimation

As described in the introduction to this guide, we are using the term *confidence interval* throughout the text to represent both *confidence interval* and *prediction interval* because of its familiarity with readers. However, in terms of a regression model, there is a technical distinction between the two terms that is of interest to actuaries estimating IBNR. To help elucidate the distinction between the two terms, it is useful to frame the terms by the following questions:

- Do we want the expected (average) value of the incurred claims PMPM corresponding to new values of the independent variables (lags 0 and 1)? We then need confidence intervals.

- Do we want the actual value of the incurred claims PMPM corresponding to new values for lags 0 and 1? We then need prediction intervals.

Note that the subtle difference in the two terms is focused on expected values versus actual values. An actuary would ordinarily be interested in the potential range of values of IBNR, which is the prediction interval. The range within which the average value of IBNR would be likely to fall is usually not of interest. We note that prediction intervals are larger than confidence intervals. Again, as earlier noted in the introduction to this guide, we will be primarily using the term *confidence interval* or *interval* throughout the text, unless further clarification is needed.

3.2.5 Interval Estimates for Simple Linear Regression

We now develop interval estimates in the context of simple linear regression. The description in this section is limited to the simple regression model.

Suppose that we want to find the relationship between a dependent variable Y and an independent variable X based on a set of sample observations collected on the two variables, given in Exhibit 3.2.

Exhibit 3.2
Sample Data for Linear Regression Example

x	y
10	8.03
8	6.90
13	7.58
9	8.81
10	8.33
14	9.96
6	7.24
4	4.26
12	10.85
7	4.82

Let $x_1,\ldots,x_n$ and $y_1, y_2,\ldots, y_n$ be these observed values (with $n = 10$).

Using simple linear regression, we will estimate the linear function that best fits this data. The model is written as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2); i = 1,\cdots,n.$$

The approach is to develop a model to predict the values of the dependent random variable Y for given values of the independent random variable X. Hence the *x*-values are considered to be fixed and the Y-values are considered to be random.

The following are the least squares estimates of the parameters:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \text{ and}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,\ i = 1,\cdots n.$$

Suppose that we want to predict the value of the dependent variable corresponding to a new observation on the independent variable X, say *x**.

The point estimate of the mean of the new observation is $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

A $100(1-\alpha)\%$ prediction interval for the new observation is

$$\left[\hat{y}^* - t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}\right)}\ ,\ \hat{y}^* + t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}\right)}\right],$$

where $t_{\alpha/2}(n-2)$ is the upper tail $\alpha/2$ value of the student t-distribution with $n-2$ degrees of freedom.

Note that the estimate is most accurate at the mean of the observed values of the independent variable $\bar{x}$ and the prediction interval is narrowest at $x^* = \bar{x}$. The further away the estimate is from $\bar{x}$, the wider the prediction interval becomes.

Using our sample data, the regression equation derived is $y = 2.885 + 0.5154\,x + \varepsilon$. Exhibits 3.3 and 3.4 illustrate output that is typically generated when a linear regression is run through a statistical package. Later on, we will discuss how some of these measures can be used to assess the appropriateness of the model.

Exhibit 3.3
Regression Output
Linear Regression Example

	Coefficient	SE Coef	t	p-Value (p)
Constant	2.8849	1.3534	2.13	0.066
X	0.5154	0.1385	3.72	0.006

Standard deviation (s) = 1.31457
R-square = 63.4%
R-square (adj) = 58.8%

The following are descriptions of the data shown in Exhibit 3.3:

- The column labeled "Coefficient" provides the estimates of the regression coefficients values in the model. The equation is found using the least-squares method that minimizes the sum of the squares of the errors.
- The "SE Coef" column provides the standard error of the estimate of the corresponding coefficient.
- The t values column provides the value of the t-statistic to test the null hypothesis that the corresponding regression coefficient is zero.
- The p-values column conveys if the data supports the corresponding null hypothesis or not. In practice, 0.05 is typically the value chosen to determine the meaning of the t-statistic, although this choice is somewhat arbitrary and may vary. If the p-value is less than 0.05, then this is viewed as the data not supporting the null hypothesis. This means that the corresponding independent variable is useful in predicting the dependent variable. Note that very small p-values are reported as zeros by most software packages. This does not mean that these p-values are zeros.
- "s" is the estimate of the model standard deviation.

- R-square (R-sq) conveys the percentage of variability in the dependent variable explained by the regression model. The R-square value has a drawback, namely, that just adding any unrelated variable into the model results in an increase in the R-square value.
- R-square (adj) (R-sq (adj)) is the R-square adjusted to remedy the shortcoming of the standard R-sq. In this sense R-sq (adj) is a good measure of how well the model explains the variability in the dependent variable.

The analysis of variance shown in Exhibit 3.4 deconstructs the total variation in the model to measure how much of it is contributed by each model component. For a linear regression model, the variation is caused by the Regression and Residual Error.

Exhibit 3.4
Analysis of Variance
Linear Regression Example

	Degrees of	Sum of	Mean		
Source	Freedom (df)	Squares (SS)	Square (MS)	F-Value (F)	p-Value (p)
Regression	1	23.932	23.932	13.849	0.006
Residual Error	8	13.825	1.728		
Total	9	37.757			

The following are descriptions of the data contained in Exhibit 3.4:

- The mean square (MS) is calculated by dividing the SS value by the corresponding degrees of freedom.
- The F-value is the value of the F-statistic used to test the overall effectiveness of the model.
- The p-value is the corresponding p-value of the test. As a general rule used in practice, if the p-value is greater than 0.05, the corresponding model is not appropriate.

To illustrate how an interval would be calculated on this sample, we will choose three points with values for x, the independent variable, of 11, 15 and 18. Note that the mean of the independent variable for our data, $\bar{x}$, is 9.30. Applying the interval formula with $\alpha = 0.05$ yields the following 95 percent prediction intervals shown in Exhibit 3.5:

Exhibit 3.5
Predicting Values
Linear Regression Example

Independent Variable	Dependent Variable	Prediction Interval		Length
11	8.554	5.329	11.780	6.451
15	10.616	6.952	14.279	7.327
18	12.162	7.939	16.384	8.445

As can be observed from Exhibit 3.5, when the difference between a given value of the independent variable and the mean of the variable increases, the length of the prediction interval also increases.

When calculating IBNR, because our main independent variable is time, increasing the difference from the mean is equivalent to projecting further out into the future. Because the width of the prediction interval is increasing as this difference increases, it is advisable to restrict regression projections to as few time periods into the future as is deemed necessary. Hence, we might use data from regression months 0 through 32 to project the PMPM values for months 33 to 35, but we would not use that same data to project the PMPM value for month 44 unless there was some particular need to do so.

3.2.6 Residuals

In regression analysis, a primary assumption is that the dependent variable is random and the independent variable is not random. Furthermore, there is a relationship between the dependent variable and the independent variables, and this relationship is fixed except for a random factor. As shown earlier, a regression equation that estimates this relationship can be derived by using a set of observations of the dependent and independent variables. This regression equation can be used to predict the value of the dependent variable for any set of values of the independent variables. The difference between the predicted value and the observed value of the dependent variable is called the *residual* for each observed value. The residuals are essentially estimates of the errors that are inherent in our model. In other words, if all the residuals were 0, the regression equation would perfectly model our data set. Given this characteristic of the residuals, an analysis of them can reveal information about the appropriateness of the model.

We test the assumption that the model is appropriate in the following ways using the residuals:

- Residual Plot
- Histogram
- Shapiro-Wilk Test

All three of these approaches are shown in the accompanying regression workbook in the Tab "Graph of Residuals."

The residual plot represents a graph of each of the residuals. If the model meets its assumptions, the graph of the residuals should not exhibit any discernable pattern, but rather appear random and unsystematic. If an expanding or contracting pattern emerges in one direction or another, then the assumption of constant variance is not supported.

The histogram is another tool used to assess the regression model. If the regression model is a good fit for the data, the histogram of residuals should exhibit a normal curve or bell-shaped, pattern. In the regression workbook, there is an option for how many bins or intervals to use in the histogram. A user can choose between 7 and 12 bins. With a histogram, the approach for assessing whether or not it is a normal curve shape is purely visual.

The other tool that can be used is a Shapiro-Wilk Test, which is a more rigorous test for normality. In this section, the residuals will be only analyzed visually.

3.3 Models Available in the Workbook

Returning to our data from Exhibit 3.1 as an example, in this section we describe the models that are available in the regression workbook accompanying this guide. Exhibit 3.6 lists the models including the applicable independent variables and regression equation:

Exhibit 3.6
Statistical Models in the Workbook

Name	Type of Model	Independent Variables	Equation
LinRegr	Linear	Time t (1)	$y = \beta_0 + \beta_1 t + \varepsilon$
QuadRegr	Quadratic	Time t	$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon$
ExpRegr	Exponential	Time t	$Y = \exp(\beta_0 + \beta_1 t + \varepsilon)$
Lin2Var	Linear	Time t_1, User Input t_2 (2)	$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \varepsilon$
Quad2VarRegr	Quadratic	Time t_1, User Input t_2	$Y = \beta_0 + \beta_1 t_1 + \beta_2 (t_1)^2 + \beta_3 t_2 + \varepsilon$
Exp2VarRegr	Exponential	Time t_1, User Input t_2	$y = \exp(\beta_0 + \beta_1 t_1 + \beta_2 t_2 + \varepsilon)$
AdjLinRegr	Linear	Time t, Weekday/weekend w_t	$y/w_t = \beta_0 + \beta_1 t + \varepsilon$
AdjQuadRegr	Quadratic	Time t, Weekday/weekend w_t	$y/w_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon$
AdjExpRegr	Exponential	Time t, Weekday/weekend w_t	$y/w_t = \exp(\beta_0 + \beta_1 t + \varepsilon)$
AdjLin2Var	Linear	Time t_1, User Input t_2, Weekday/weekend w_t	$y/w_t = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \varepsilon$
AdjQuad2VarRegr	Quadratic	Time t_1, User Input t_2, Weekday/weekend w_t	$y/w_t = \beta_0 + \beta_1 t_1 + \beta_2 (t_1)^2 + \beta_3 t_2 + \varepsilon$
AdjExp2VarRegr	Exponential	Time t_1, User Input t_2, Weekday/weekend w_t	$y/w_t = \exp(\beta_0 + \beta_1 t_1 + \beta_2 t_2 + \varepsilon)$

We note the following:

(1) The values of Time (t or t_1) are the positive integers assigned to each month.

(2) The input variable (t_2) is an unspecified variable whose values for each month are assigned by the user. In the initial example in this section, the use of this variable is limited to values of 0 and 1. This variable enables the user to differentiate certain months from the others. For example, an actuary may have information about hospital pricing changes that occurred at a certain date. In this instance, it may be appropriate to model the data before and after the pricing change by applying this variable.

3.3.1 Key Steps to Applying the Methodology Discussed Above to Our Data

There are a number of approaches for determining the most appropriate regression model. For the purposes of this section of the document, it is recommended that the overall approach focus on maximizing adjusted R-square for all of the models that are tested. Other approaches may use the p-values of the coefficients and other diagnostics to help determine the most appropriate model. Readers may wish to consult *Econometric Models and Economic Forecasts* by Pindyck and Rubinfeld to learn more about regression approaches.

The following is a high-level outline of the steps for determining estimates for lags 0 and 1. The steps describe an iterative approach for selecting an appropriate model. An alternative is to simply run all models and compare the adjusted R-square values. In either case, the goal for this section of the document is to maximize adjusted R-square. We will discuss how these steps are applied in terms of the regression workbook in Section 3.4.

1. Plot the dependent variable against the independent variables to observe the visual patterns of the data.

2. Based on a visual inspection of the plot, determine which model programmed in the regression workbook appears to be an appropriate choice for our data. In addition to the model that the user thinks is most appropriate, it is recommended to run the simple regression model as a baseline for comparison against other models.

3. Fit the model and review the diagnostics. The key statistical value to review is the adjusted R-square value, with the goal of maximizing it. It is also useful to review the residuals, checking the normality and constant variance assumptions, and spotting outliers.

   In addition to assessing the model for statistical appropriateness, the model should be reasonable based on knowledge of outside factors affecting the health care plan or book of business for which incurred claims are being estimated.

4. For comparison purposes, run additional models as necessary and review the diagnostics. When trying different models, it is useful to change only one model component with each successive model that is run. The benefit of this approach is that it is easier to gauge the impact of the addition or deletion of each individual model component. For instance, say the first model run is a linear regression model with one independent variable—time. If you want to improve on this model, you might next add either the weekend/weekday variable or the additional time variable, but not both at the same time.

3.4 Using the Regression Workbook

This section outlines the steps for deriving the IBNR estimates through use of the regression workbook.

1. Open up the file and navigate to Tab "OriginalData." The statement and valuation dates are fixed at December 31, 2003. Column A lists the incurred months and Row 4 lists the paid months. The claims data is shown in tabular form. In Column AO, the values for the second variable are input.

2. Calculate IBNR for all months by pressing "Alt-F8," selecting the macro "CalcCompletionFactor" and selecting "Run." In order to run the macros, you may need to adjust the macro security level in Excel. You will be asked: "How many months are in the latest year of your triangle?" For the sample data, the answer is 12. After the macro "CalcCompletionFactor" is completed, the Tab "Graph of PMPM" graphically represents the incurred claims estimates for months 0 through 35 derived using the completion factor method. (If you do not want to use the completion factor method for the two most recent months, ignore the values shown for months 34 and 35.)

3. If you plan on using Model Lin2Var (one of the model choices listed in Exhibit 3.6) or any other model that has "2Var" in its name, set the values in Column AO in Tab "Original Data" to the desired values. Column AO represents the User Input t_2. For the purposes of this section, we limit modelling of this variable to the following conditions:
    - It only has values of 0 or 1.
    - It can only change between 0 and 1 once. In other words, we will set the value of this variable to 0 for lag 0. If we set the value of this variable to 1 for lag n ($n > 0$), all lags greater than n will have this variable set to 1.

4. If we want to apply a different method for estimating months 34 and 35 (which usually will be the case), we select "Alt-F8" again and select the macro whose name corresponds to the model listed in Exhibit 3.6 that we want to use to derive PMPM estimates for lags 0 and 1. After selecting the desired model, you will be asked: "How many months do you want to include in this regression?" If you are using the complete set of data points, the answer for our example is 34, corresponding to the months January 2001 through October 2003. If there is a reason for eliminating the data point January 2001, input 33, corresponding to the months February 2001 through October 2003. The number of months selected will always correspond to the same number of most recent months in the regression. The minimum number of data points that must be selected is 10.

    At this point, the user is then prompted to input a confidence interval.

Two clarifications are needed for this input:

- A confidence level needs to be input in order to calculate an interval, regardless of whether it is a confidence interval or a prediction interval. The type of interval depends on the estimate that is desired (an average value or a single value). The magnitude of the confidence level depends on how certain we want to be that the estimated value falls within the calculated (confidence or prediction) interval.

- When the desired confidence level is input, ensure that the decimal separator corresponds to your computer's language settings. For most anticipated users, it will be a period, but for some it may be a comma. If the wrong decimal separator is used, the macro will fail.

5. After the desired confidence level is input, the macro completes and the cursor lands on the Tab "Graph of Residuals," where key output is shown.

6. In Tab "Graph of Residuals," the following graphs and statistics appear:
    - Graph of residuals
    - Normal probability plot
    - Histogram of residuals
    - Shapiro-Wilk Test for normality of residuals showing both the Shapiro-Wilk W statistic and the p-value range.

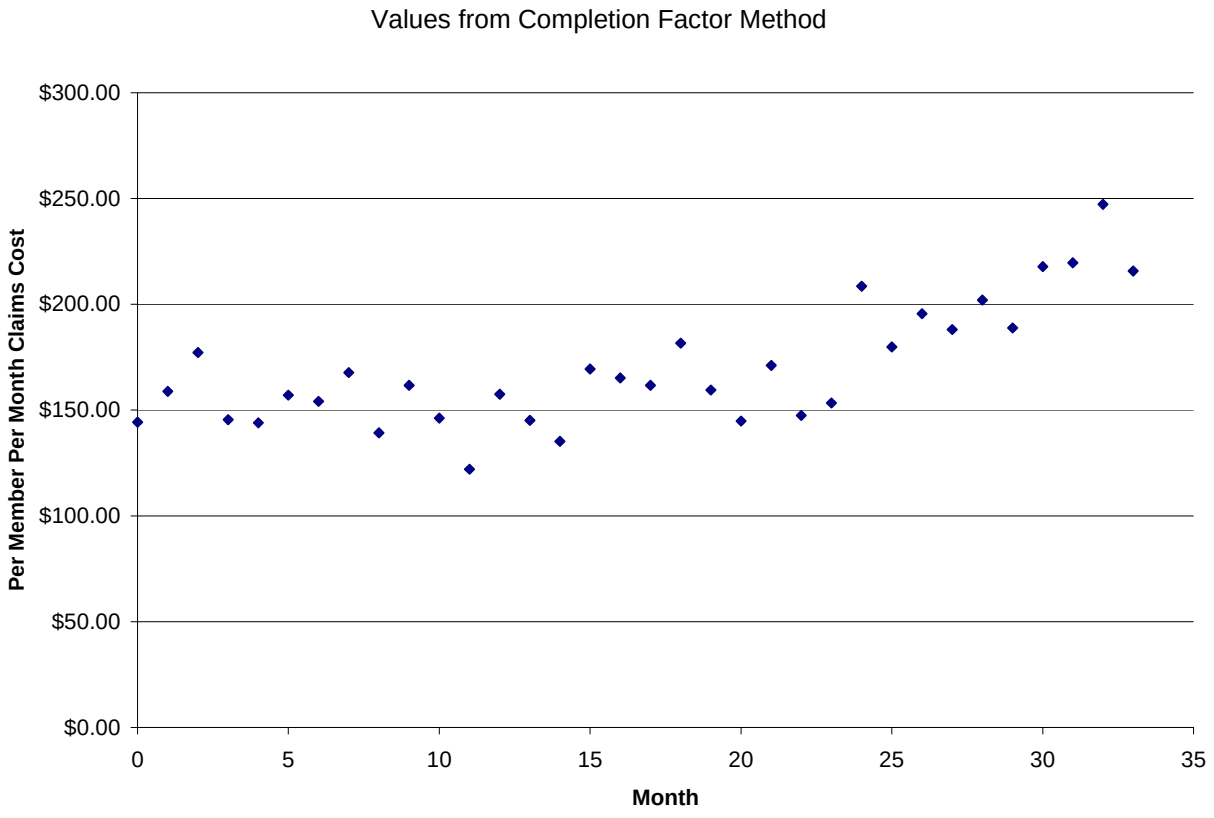
7. In the Tab "Recast Example," a recasting calculation is illustrated. Recasting is when the IBNR calculation is compared with the actual claim runout experience. For instance, consider a best estimate IBNR calculated in October 2003. Two months later, in December 2003, with an additional two months of paid claims, recasting would compare the original October 2003 IBNR estimate with an updated IBNR estimate for October 2003 incorporating the additional data.
8. To use the workbook with different data for membership, claims and weekday/weekend factors, input this data in the following areas of the workbook:
    - Membership data in Column B in Tab "Original Data"
    - Claims data starting in Cell D6 of Tab "Original Data"
    - Weekday/weekend factors in Tab "Weight factors"

    Note: The workbook is designed to use exactly 36 months of claims data. If less than 36 months is input, the program will not run correctly. If less than 36 months of actual claims data is available, a workaround for this limitation is to estimate claims data for any missing months.

3.5 Numerical Example: Deriving IBNR for Our Data Set

As a first step, we graph and visually inspect our data. The plot of the 34 PMPM estimates for the time period January 2001 through October 2003 exhibits the following pattern:

Exhibit 3.7
Graphical View of the Data for the Regression Analysis Example



Instead of a pattern of constantly increasing claims costs over time, the graph seems to exhibit two relative levels of claims costs, with one level for months 0 through 23 and another level for months 24 through 33; therefore, we will assign the User Input t_2 (0,1) variable as follows:

- 0 for months 0 through 23;
- 1 for months 24 though 35.

Note that we are assuming that the claims pattern that commenced at lag 24 will continue on for lags 34 and 35, the two lags that we are predicting.

Based on this observation, the first two models we choose to run for this example are the following:

- LinRegr: Linear regression model as a base;
- Lin2Var: Linear regression model with the User Input t_2 variable defined on the basis of the pattern observed in Exhibit 3.7 previously described.

The key output values from LinRegr from the "Regression-Results" Tab in the regression workbook are shown in Exhibit 3.8:

Exhibit 3.8
Regression Results
LinRegr Model

Estimators for coefficients		**p-values**
beta_0	135.2215	7.03E-20
beta_1	2.095960	9.19E-07
Standard Error	19.79333	
R^2	0.534173	
Adjusted R^2	0.519616	
p-value for F-test	9.19E-07	

This model has an adjusted R-square of 0.520. For the purposes of our estimate, we wish to see if this value can be improved. Therefore, if we cannot improve on our adjusted R-square value, we would use LinRegr.

The key output values from Lin2Var are shown in Exhibit 3.9:

Exhibit 3.9
Regression Results
Lin2Var Model

Estimators for coefficients		**p-values**
beta_0	148.5258	7.83E-22
beta_1	0.526894	2.49E-01
beta_2	42.78974	1.07E-04
Standard Error	15.72873	
R^2	0.715039	
Adjusted R^2	0.696654	
p-value for F-test	3.54E-09	

With Lin2Var, our adjusted R-square has improved from 0.520 to 0.697, an increase of 0.177. This relatively high increase in adjusted R-square indicates that our second time variable (referred to as a "Boolean variable") is contributing significantly in predicting the value of the dependent variable.

Further assessing the best model for our data, we add the Weekday/weekend w_t variable, which has the effect of transforming the Lin2Var model into AdjLin2Var. The results of the AdjLin2Var model are shown in Exhibit 3.10:

Exhibit 3.10
Regression Results
AdjLin2Var Model

Estimators for coefficients		**p-values**
beta_0	153.5872	1.39E-23
beta_1	0.523993	2.05E-01
beta_2	43.62881	2.09E-05
Standard Error	14.20020	
R^2	0.760245	
Adjusted R^2	0.744777	
p-value for F-test	2.43E-10	

The adjusted R-square value for the AdjLin2Var model is 0.745, which is an improvement over the adjusted R-square value of 0.697 that was obtained from the Lin2Var model. Thus, the addition of the Weekday/weekend w_t variable has improved the model's predictive power.

Continuing on with our pursuit of the best model, we now test to see if the exponential model, AdjExp2VarRegr, results in an improved R-square value. Recall from Exhibit 3.6 that the AdjExp2VarRegr model includes the User Input t_2 and Weekday/weekend w_t variables. The results of applying this model are illustrated in Exhibit 3.11.

Exhibit 3.11
Regression Results
AdjExp2VarRegr Model

Estimators for coefficients		**p-values**
beta_0	5.036368	2.66E-47
beta_1	0.002916	2.02E-01
beta_2	0.234549	3.12E-05
Standard Error	0.078531	
R^2	0.752189	
Adjusted R^2	0.736201	
p-value for F-test	4.06E-10	

With this model, our adjusted R-square value decreases to 0.736 from the adjusted R-square value of 0.745. As a result, we will not use this model because it is more complicated than the previous models and it yields a lower adjusted R-square.

With the conclusion that the exponential model, AdjExp2VarRegr, does not improve the adjusted R-square, we now try the quadratic model, AdjQuad2VarRegr. This model includes the User Input t_2 and Weekday/weekend w_t variables. The results of using the AdjQuad2VarRegr model are shown in Exhibit 3.12.

Exhibit 3.12
Regression Results
AdjQuad2VarRegr Model

Estimators for coefficients		**p-values**
beta_0	163.0751	1.09E-21
beta_1	-1.646172	1.07E-01
beta_2	0.085859	2.46E-02
beta_3	25.54311	2.93E-02
Standard Error	13.25123	
R^2	0.797954	
Adjusted R^2	0.777749	
p-value for F-test	1.54E-10	

As can be observed from Exhibit 3.12, the adjusted R-square value with this model is 0.778, which is the largest adjusted R-square value we have obtained. We select this model for further analysis.

In certain cases, the introduction of additional variables into the model does not increase its accuracy. This effect is usually referred to as multicolinearity. This topic is discussed in more depth in the appendix.

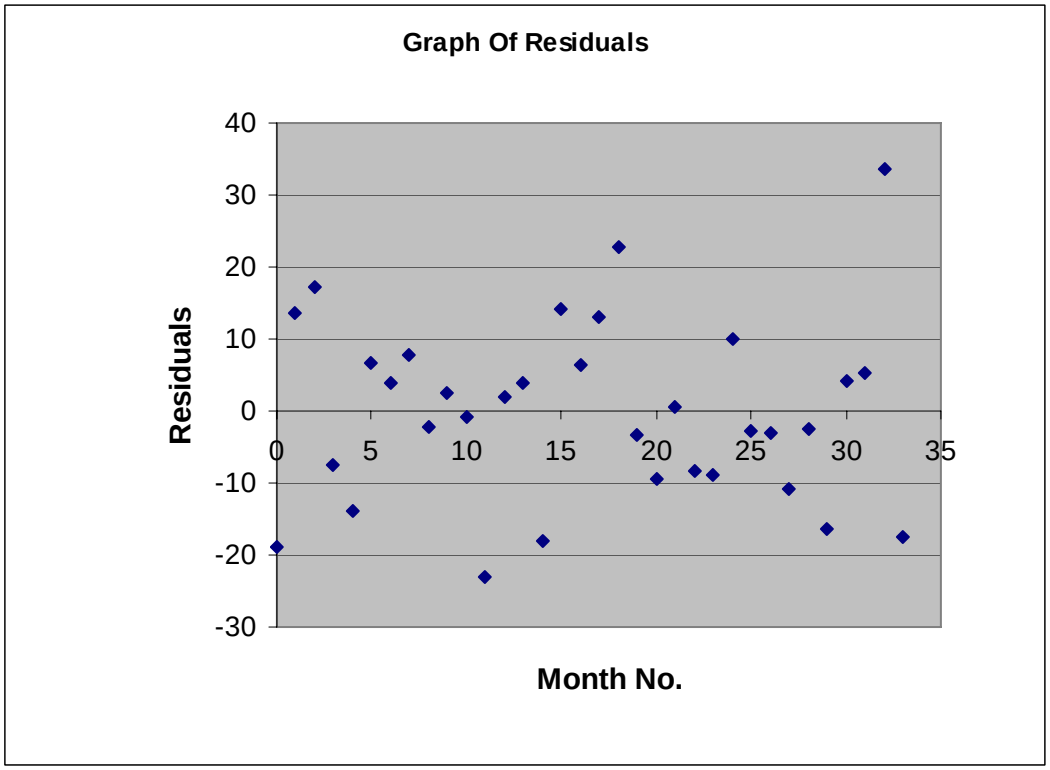
As a check for reasonableness of the model choice, we compare the adjusted R-square value of all of the models included in the regression workbook. Exhibit 3.13 lists the models and corresponding adjusted R-square value.

Exhibit 3.13
Summary of Adjusted R-Square Values

Name	Type of Model	Adjusted R-Square Value
LinRegr	Linear	0.519616
QuadRegr	Quadratic	0.721356
ExpRegr	Exponential	0.502923
Lin2Var	Linear	0.696654
Quad2VarRegr	Quadratic	0.739865
Exp2VarRegr	Exponential	0.675682
AdjLinRegr	Linear	0.552597
AdjQuadRegr	Quadratic	0.747338
AdjExpRegr	Exponential	0.548943
AdjLin2Var	Linear	0.744777
AdjQuad2VarRegr	Quadratic	0.777749
AdjExp2VarRegr	Exponential	0.736201

As this exhibit shows, the AdjQuad2VarRegr model has the highest adjusted R-square value. The residual graph for the AdjQuad2VarRegr model appears in Exhibit 3.14.

Exhibit 3.14
Residuals
AdjQuad2VarRegr



Note that the residuals for the AdjQuad2VarRegr model do not exhibit a consistent visual pattern and appear to be random. If the graph of the residuals did exhibit a pattern, it would indicate that the model is not meeting the regression assumptions. Therefore, Exhibit 3.14 is confirmation of the appropriateness of the model based on an examination of the residuals.

Using the AdjQuad2VarRegr model results in PMPM incurred value estimates of $207.70 for November 2003 and $230.08 for December 2003. Before accepting the values from the AdjQuad2VarRegr model as our estimates, one more reasonability check is in order. Comparing the PMPM values from the completion method, the figures for January 2003 to October 2003 are shown in Exhibit 3.15.

Exhibit 3.15
Testing for Reasonableness of Results
2003 Per Member Per Month Input Values

Jan-03	$208.59
Feb-03	179.89
Mar-03	195.55
Apr-03	188.06
May-03	202.00
Jun-03	188.86
Jul-03	217.83
Aug-03	219.59
Sep-03	247.26
Oct-03	215.69

Scanning Exhibit 3.15 for the most recent year's values, the PMPM values of $207.70 for November 2003 and $230.08 for December 2003 appear to be within a reasonable range of values we would have expected. For instance, if the values were, say, $600 or $20, we would reject the estimates as being either too high or too low. However, since the values for November and December 2003 appear consistent with other months, we accept the values from AdjQuad2VarRegr as our PMPM estimates.

3.5.1 Adjustment for Catastrophic Cases

Recall that for the completion factor method calculation, a $750,000 catastrophic case was removed (refer to Section 2.2 for a general discussion of outliers). Consequently, the IBNR calculation presently does not include any reserve for catastrophic cases. In practice, there are several ways to treat catastrophic claims for IBNR purposes. Among possible approaches would be to include it in the experience in the process of developing completion, but to adjust it to a pre-determined amount consistent with an internal pooling arrangement, external reinsurance agreement, or other reasonably expected amount. Another approach is to simply remove it from the experience and set up a separate catastrophic reserve for catastrophic claims.

In preparing financial statements, it is important to document the approach for catastrophic case reserving as well as describing any risk of understatement due to catastrophic cases, if needed. From a practical perspective, the best way to minimize the risk of an understatement due to catastrophic claims is to obtain relevant information about them from case management or other internal company sources as close as possible to their original incurral date. In other words, although the occurrence of catastrophic claims cannot be minimized, the lag time for properly accounting for them can be minimized.

3.6 Calculating IBNR

In Section 3.5, we derived PMPM incurred claims estimates for November and December 2003. To calculate IBNR from these estimates, the following formula is used:

IBNR = Incurred Claims – Claims paid to date

where Incurred Claims = PMPM × Membership.

The IBNR calculation for November and December 2003 is illustrated in Exhibit 3.16:

Exhibit 3.16
Calculation of IBNR for Lags 0 and 1

	Incurred Month		
	November 2003	December 2003	
PMPM Estimate	$207.70	$230.08	(1)
Members	11,902	11,844	(2)
Total Incurred Claims	$2,472,045	$2,725,068	(3) = (1) * (2)
Claims Paid			
Paid Month			
Nov 2003	$58,510	N/A	(4) *
Dec 2003	$1,225,307	$96,378	(5)
Total Paid	$1,283,817	$96,378	(6) = (4) + (5)
IBNR	$1,188,228	$2,628,690	(7) = (3) – (6)

* By definition, in our data, the paid date occurs on or after the incurred date; therefore, the value for claims incurred in December 2003 and paid in November 2003 is not applicable. The value $58,510, claims paid and incurred in November 2003, can be found in Cell AL40 in Tab "OriginalData" in the "Statistical Methods – Regression Method – Workbook.xls."

Our total IBNR is shown in Exhibit 3.17.

Exhibit 3.17
Total IBNR as of December 2003

December 2003 – Lag 0	$2,628,690	(1)
November 2003 – Lag 1	1,188,228	(2)
December 2002 Through October 2003 – Lags 2 through 12	833,796	(3)
Sub-total	4,650,714	(4)=(1)+(2)+(3)
IBNR for Catastrophic Cases	218,750	(5)
Total	$4,869,464	(6)=(4)+(5)

For the purposes of this calculation, we are assuming a separate reserve for catastrophic cases of $218,750. Recall that a catastrophic claim of $750,000 was removed from the data. To account for catastrophic claim exposure and the removal of the $750,000 claim, $218,750 has been added into the calculation of the Total IBNR. As noted in Section 3.5.1, there are a number of approaches for estimating IBNR due to catastrophic claims. Discussion of these approaches is beyond the scope of this document. It should be noted that depending on the ultimate catastrophic claim amount experienced, the $218,750 estimate may cause a situation of under- or over-reserving. Furthermore, the variability in the frequency and severity of catastrophic claims often represents a source of added complexity in determining the final IBNR for accounting purposes.

3.7 Confidence Intervals

The upper bound of the 95 percent confidence interval is shown in Exhibit 3.18. The interval length results are obtained from the Tab "Regression Results" of the regression workbook.

Exhibit 3.18
Calculation of the Upper Bound of a Prediction Interval (with 95% Confidence Level)

	November 2003	December 2003	Total	
PMPM Value	$207.70	$230.08		(1)
One-Sided Prediction Interval Length	28.82	33.00		(2)
Upper Bound—Prediction Interval	236.52	263.08		(3) = (1) + (2)
Members	11,902	11,844		(4)
One-Sided Interval for All Members (Margin)	$343,016	$390,852	$733,868	(5) = (2) * (4)
IBNR—Best Estimate			$4,650,714	(6) Line (4) – Exhibit 3.17
Upper Bound of IBNR with 95% Prediction Interval			$5,384,582	(7) = (5) + (6)
Margin as Percent of Best Estimate			15.8%	(8) = (5) ÷ (6)

Applying the results from Exhibit 3.18, we are 97.5 percent confident (the upper bound is only one side of the confidence interval) that, after all of the claims incurred on or before December 2003 have been paid, the total paid claim amount will be less than or equal to $5,384,582, excluding catastrophic claims. The margin (item 8) is the amount added to have sufficient IBNR to cover paid claims 97.5 percent of the time.

Note that there was no confidence interval assigned around the catastrophic case IBNR.

If different levels of confidence are desired for the prediction intervals, we can use the regression workbook to derive these levels. Other prediction levels with the AdjQuad2VarRegr model are shown in Exhibit 3.19. These values were obtained by running the model with different prediction intervals.

Exhibit 3.19
IBNR with Different Prediction Intervals

Prediction Interval	One Sided Prediction Interval	IBNR With Margin (Without Catastrophic IBNR)	IBNR With Margin (With $218,750 Catastrophic IBNR)
95%	97.5%	$5,384,582	$5,603,332
50%	75%	4,896,088	5,114,838
90%	95%	5,260,648	5,479,398
99%	99.5%	5,638,978	5,857,728

A major limitation of the method described in this section for calculating prediction intervals is that the intervals are just based on our calculations for the most recent two months (lags 0 and 1). In Section 4, we discuss methods for calculating prediction intervals for all lags.

3.8 Nonuniform Data

In previous sections, we assumed that the underlying benefit plan resulted in uniform data. In this section, we examine adjustments to the regression models when the pattern of data is impacted by a change in the benefit plan or introduction of a high deductible plan.

3.8.1 Scenario 1: Known Benefit Changes

Common health plan benefit changes include increasing deductibles and co-pays as well as reductions in covered services. Consider a health plan that wishes to reduce benefits by 5 percent starting January 1, 2002 and an additional 10 percent reduction effective January 1, 2003 through a combination of benefit changes.

Exhibit 3.20 illustrates the relative value of plan benefits for this health plan for years 2001–2003, assuming a factor of 1.000 before any changes.

Exhibit 3.20
Scenario 1—Known Benefit Changes
Relative Value of Plan Benefits

Year	Factor
2001	1.000
2002	0.950
2003	0.855

What is an appropriate way to model the benefit changes in the regression model and estimate the IBNR for the most recent lag months?

Since we will be estimating values for 2003 and onward using data from 2001 and 2002, we will want all the regression model data to reflect the impact of the benefit changes that were applied

through the end of 2003. In other words, we want the value of the benefit plan to be normalized (sometimes also referred to as "on level") for all years of data in the regression model. A similar adjustment approach is also typically applied when estimating future claim costs for pricing purposes.

To normalize the regression model data, the adjustment factors shown in Exhibit 3.21 would be multiplied by incurred claim cost data for the corresponding incurred year.

Exhibit 3.21
Scenario 1—Known Benefit Changes
Regression Data Adjustment Factors
By Incurred Year

Incurred Year	Factor
2001	0.855
2002	0.900
2003	1.000

The logic for the adjustment factors is as follows:

- For 2003, the incurred claims data reflects all benefit plan changes, so no adjustment is needed. Therefore, the factor is 1.000.

- For 2002, the incurred claims data reflects the 5 percent decrease on January 1, 2002, but not the 10 percent decrease on January 1, 2003. Therefore, the 2002 data needs to be adjusted by a factor of 0.900 to represent a 10 percent reduction.

- For 2001, the incurred claims data reflects neither the January 1, 2002 nor January 1, 2003 reductions. Therefore, the 2001 data needs to be adjusted by the compound effect of the 10 percent and 5 percent reduction: 0.90*0.95 = 0.855.

Once the data has been adjusted, the appropriate regression model can be selected using the process described earlier in this section, focusing on maximizing adjusted R-square.

3.8.2 Scenario 2: High Deductible Health Plan

In Scenario 1, known benefit changes were modelled. For Scenario 2, it will be assumed that a new high deductible health plan is instituted effective January 1, 2002. This high deductible plan significantly changes the pattern of claims over the course of a calendar year. Scenario 2 is more complicated than Scenario 1 because the new benefit plan seasonally affects the claim pattern.

With a high deductible plan, claim activity is reduced significantly in the early part of the year as insureds are accumulating costs towards the deductible. Exhibit 3.22 illustrates the relative value of the benefits by month before and after the introduction of the high deductible plan for Scenario 2. We will assume claim costs have a relative factor of 1.000 before introduction of the high deductible plan.

Exhibit 3.22
Relative Factors Used to Model High Deductible Plan

Month	Before High Deductible Plan - 2001	After High Deductible Plan - 2002
January	1.000	0.100
February	1.000	0.250
March	1.000	0.500
April	1.000	0.750
May through December	1.000	1.000

What is an appropriate way to model the new benefit plan in the regression model and estimate the IBNR for the most recent lag months?

We will describe two approaches that can be used to model this.

- Approach 1

The first approach follows a similar logic to that used in Scenario 1, except with a slight alteration. Rather than adjust the data to reflect the impact of the change after the high deductible plan was put into effect, we will normalize the data to reflect the plan value before the high deductible plan was put into effect. We adjust the data in this way because the high deductible plan data exhibits a more complicated claim pattern than the previous plan, making it more difficult to model.

Extending this logic, for 2001 incurred data, no adjustment is made. For 2002 incurred data, adjustments are made to back out the impact of the high deductible plan. For example, a claim incurred in January 2002 would be divided by 0.100 to normalize it to the 2001 data. Similarly, a claim incurred in March 2002 would be divided by 0.500.

After the data has been adjusted, the appropriate regression model would be selected based on the process described earlier. Once the model has been selected, estimates for the IBNR can be calculated. However, a final seasonality adjustment must be made to the estimates produced by the regression model because the model does not reflect the impact of the high deductible. For example, an IBNR estimate produced by the model for February 2003 would be multiplied by 0.250 to account for the seasonality of the high deductible plan.

- Approach 2

Rather than adjust the actual data as in Approach 1, we could use the models with the w_t factor to model the high deductible plan. Inputting the w_t column in the Tab "Regression" with the After High Deductible Plan relative factors values shown in Exhibit 3.22 will accomplish the same adjustment as in Approach 1. Once the regression model is selected and estimates produced by the model, a similar final adjustment for seasonality would be made to those estimates.

3.9 Summary

In this section, a regression method was used to calculate IBNR for the most recent two months combined with the completion factor method for prior months. Using the regression methodology enabled us to establish a confidence level for the IBNR calculated in the last two months. Since 70 percent of IBNR is typically concentrated in the last two months, a confidence interval for these last two months is a good initial estimate of the variance in the total IBNR estimate. This section also illustrated how the regression method could be applied in instances of benefit plan changes.

In the next section, we will introduce a method for calculating the confidence interval for the entire IBNR through simulation.

Section 4—Simulation Techniques to Estimate Confidence Intervals for IBNR Reserves

4.1 Introduction to Simulation

In previous sections, regression was used to calculate confidence intervals for the IBNR estimates of the most recent months of claim data. This section introduces a different approach—simulation (or sometimes formally referred to as Monte Carlo simulation). The techniques described in this section were suggested by Walter James, ASA, MAAA, life & health actuary for the North Carolina Department of Insurance. He uses simulation as a practical way to convey the level of variability in his IBNR estimates for different audiences.

In general, simulation is an approach through which certain values in a spreadsheet or other program are recalculated hundreds of times, with the goal of describing a distribution of possible outcomes. For our purposes, the possible outcomes that we will focus on are the values of the total IBNR for ***all*** months in our data (rather than just the most recent months). By using simulation, we can describe a distribution of possible values of the total IBNR and use this distribution to calculate various confidence intervals as well as other statistics.

Another key element of simulation is generating random values repeatedly for the input variables that feed the IBNR calculation. A new set of random values is generated each time the computer calculates a possible outcome. One cycle of this process is sometimes referred to as a model "iteration." In previous sections, the input variables feeding the IBNR calculation have been completion factors or PMPM factors. For the example illustrated in this section, we will use the average claim amount per member as the primary input variable for the IBNR calculation. However, the simulation technique may also be applied to completion factors, completion ratios, or other types of PMPM factors when they are used as the primary input variables in the IBNR calculation.

In describing the input variables, we are using the term "random values" somewhat loosely. They are not completely random. Instead, they are generated from one or more probability distributions that we will use to describe the input variables. How can we determine an appropriate probability distribution to describe the input variables? The process that we will use is known as distribution fitting. In simple terms, we will try several types of probability distributions with specific parameters and then use a statistical test to rate them and select the most appropriate.

To compare the probability distributions and rate them, there are several types of statistical tests that can be employed. For the illustration in this section, a Chi-square test is used to rate the distributions. Further information about Chi-square tests can be found in statistical textbooks or reference pages on the Internet. Once the appropriate probability distributions have been selected to describe the input variables, the actual simulation process begins with values for the input variables generated repeatedly and the resulting output variable (the total IBNR) recorded for each iteration.

To illustrate the approach described in the preceding paragraphs, we will use a simple example and outline each step on the following pages. To perform the distribution fitting and simulation calculations, we will use an Excel add-in program called @RISK. (A trial version of @RISK can be obtained from www.palisade.com.) For the example illustrated in this section, either the Professional or Industrial version of @RISK is needed because each contains distribution fitting capability (the Standard version of @RISK does not). Other simulation software packages are available for purchase as well as several freeware packages that can be downloaded from a number of different Web sites. For instance, another Excel add-in for purchase that performs simulation is Crystal Ball, which can be obtained from www.crystalball.com.

4.2 Illustrative Example

An Excel workbook containing the illustrative example can be downloaded from the same Web page on the SOA Web site on which this guide is available. The name of the workbook file is "Simulation Example.xls." It would be helpful to have this workbook open as you follow these steps. Also, to complete certain steps in Section 4.3, @RISK should be launched and running in Excel.

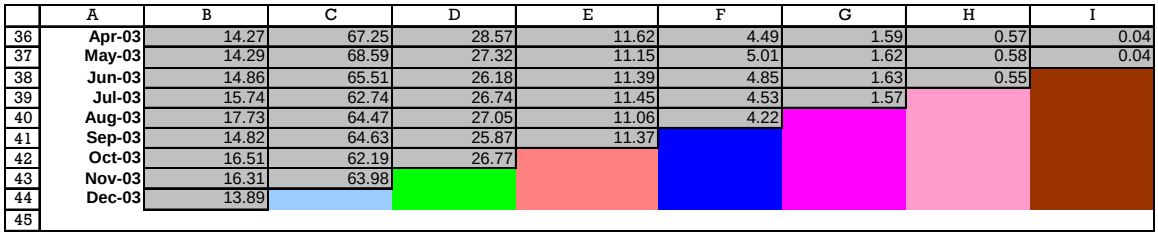
The workbook contains two spreadsheets: "Claims and Membership Data" and "Claim PMPM and Calculations."

The "Claims and Membership Data" spreadsheet contains the source data for the IBNR calculation. For illustrative purposes, the data has been truncated to show history of claims paid for a given incurred month for up to seven months following that incurral month. Consequently, for this example, we will assume that all claims are paid by the seventh month following the incurred month. Membership for each incurred month also appears on this spreadsheet.

The "Claim PMPM and Calculations" spreadsheet converts the claim amounts for each month in the grid that appears in the "Claims and Membership Data" spreadsheet into per member per month (PMPM) amounts.

In each column in this spreadsheet, there is a series of colored cells that appear at the bottom of the column (see Figure 4.1). These cells represent paid months for a particular incurred month that have not yet occurred and are values that we will be estimating. They also represent the input variables for the simulation we will perform. Essentially, we will specify a probability distribution to describe each of the columns and then generate values for the colored cells based on the applicable probability distribution.

Figure 4.1
Outstanding Payment Months (Colored Cells)
Prior to Distribution Fitting and Simulation



	A	B	C	D	E	F	G	H	I
36	Apr-03	14.27	67.25	28.57	11.62	4.49	1.59	0.57	0.04
37	May-03	14.29	68.59	27.32	11.15	5.01	1.62	0.58	0.04
38	Jun-03	14.86	65.51	26.18	11.39	4.85	1.63	0.55	
39	Jul-03	15.74	62.74	26.74	11.45	4.53	1.57		
40	Aug-03	17.73	64.47	27.05	11.06	4.22			
41	Sep-03	14.82	64.63	25.87	11.37				
42	Oct-03	16.51	62.19	26.77					
43	Nov-03	16.31	63.98						
44	Dec-03	13.89							
45									

To the right of the claims data is a small table that contains the IBNR calculations (see Figure 4.2). We will sum all PMPM amounts for the outstanding paid months for a particular incurred month and multiply this by the membership for that incurred month. This calculation will be done for all incurred months that still have outstanding payments, i.e., seven months have not yet elapsed since the incurred month. The sum of all months then represents the Total IBNR (our output variable). When the worksheet is first opened prior to running the simulation, columns K and M in the IBNR calculations table will show $0, as in Figure 4.2. The table appears as such because values for the formulas contained in columns K and M in the IBNR calculations table have not yet been generated by the simulation.

Figure 4.2
The IBNR Calculations Table prior to Simulation

	K	L	M
34	**IBNR Calculations**		
35	**Sum of PMPM**		
36	**for Outstanding**		
37	**Paid Months**	**Membership**	**IBNR**
38	$0	226,398	$0
39	$0	226,590	$0
40	$0	225,432	$0
41	$0	226,333	$0
42	$0	225,886	$0
43	$0	226,742	$0
44	$0	226,809	$0
45		**Total IBNR**	**$0**

The process of generating values for each of the colored cells will occur over and over again depending on the number of iterations we specify (for our illustrative example, we will use 10,000). As a result, a new set of values for the cells in Columns K and M will appear in the IBNR calculations table during each iteration. At the same time, a new Total IBNR will be calculated and stored by @RISK for each iteration to form a sample probability distribution of the Total IBNR.

Again, we should emphasize that this is a simplified example. The complexity of the IBNR calculation can be greatly enhanced depending on the data that is available and the method that is applied. Our intent is to illustrate how simulation can be used to specify statistics that describe the IBNR.

The following are steps to complete the simulation.

4.3 Step by Step Guide

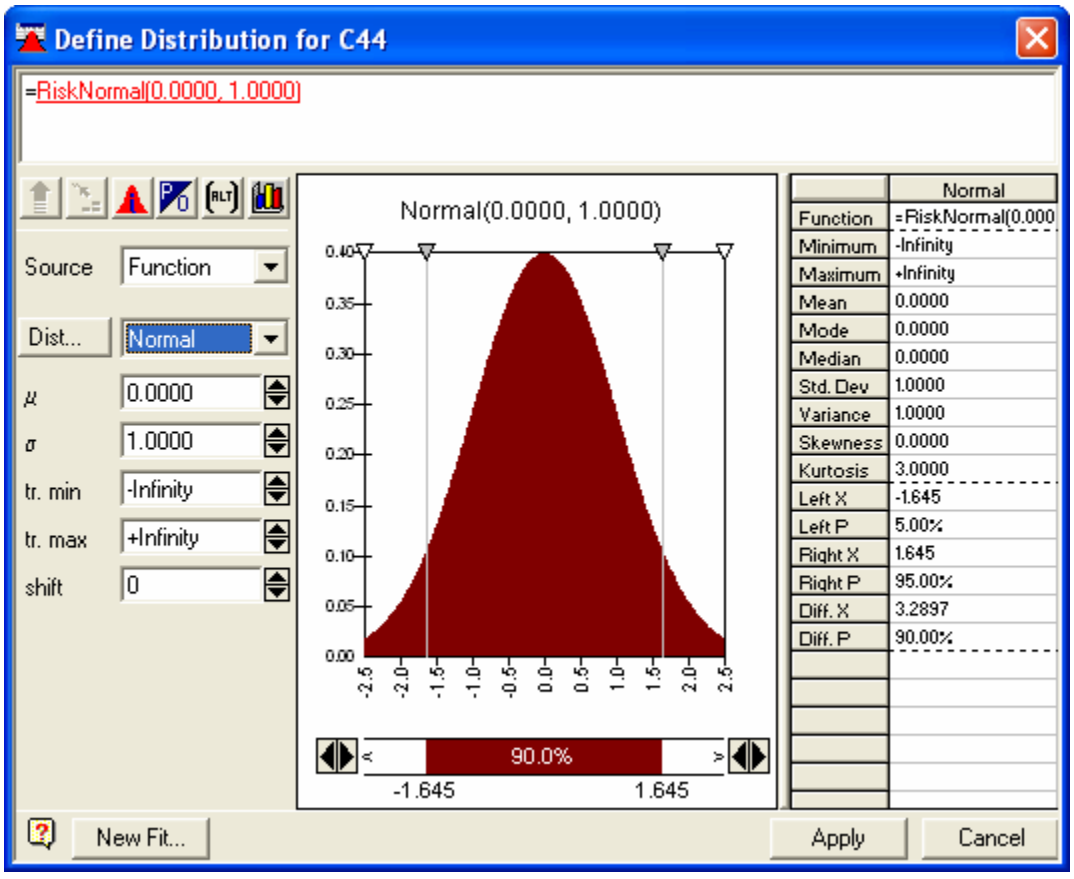
Step 1—Place the cursor on Cell C44 (shaded in blue).

Step 2—Right click the mouse on Cell C44. A window will appear with **@RISK** as one of the choices.

Step 3—Move to the @RISK choice and select **Define Distribution** from the @RISK subwindow that appears.

Step 4—A window titled **Define Distribution for C44** will appear (see Figure 4.3).

Figure 4.3
The Define Distribution Window prior to the Distribution for Cell C44 Being Fitted



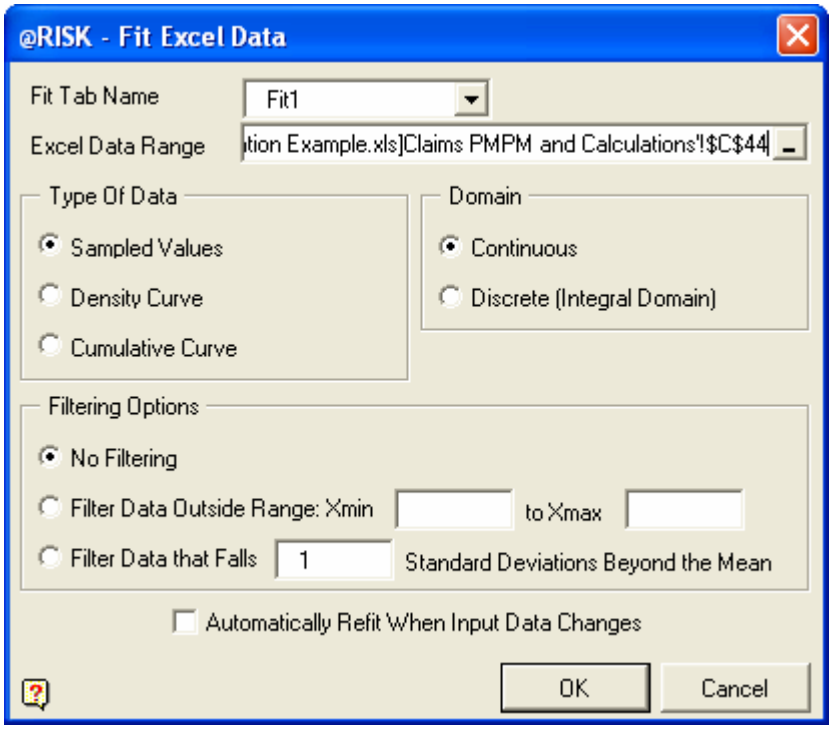
Step 5—Click on the **New Fit** button in the lower left hand corner of the window.

Step 6—A new window titled **Fit Excel Data** will appear (see Figure 4.4). The selections on this window tell @RISK which data is being used to fit the probability distribution. The window has options for specifying the Type of Data, Domain and Filtering Options. The filtering options can be used if there are outliers present in the claim or membership data.

Step 7—Specify the **Excel Data Range** in the **Fit Excel Data** window by highlighting the applicable cells to be fitted. For Cell C44, the applicable cells are C9 through C43. For Cells D43 and D44, applicable cells are D9 through D42.

Step 8—Click **OK** in the **Fit Excel Data** window.

Figure 4.4
The Fit Excel Data Window prior to Specifying the Range of Data
to Be Used for the Simulation

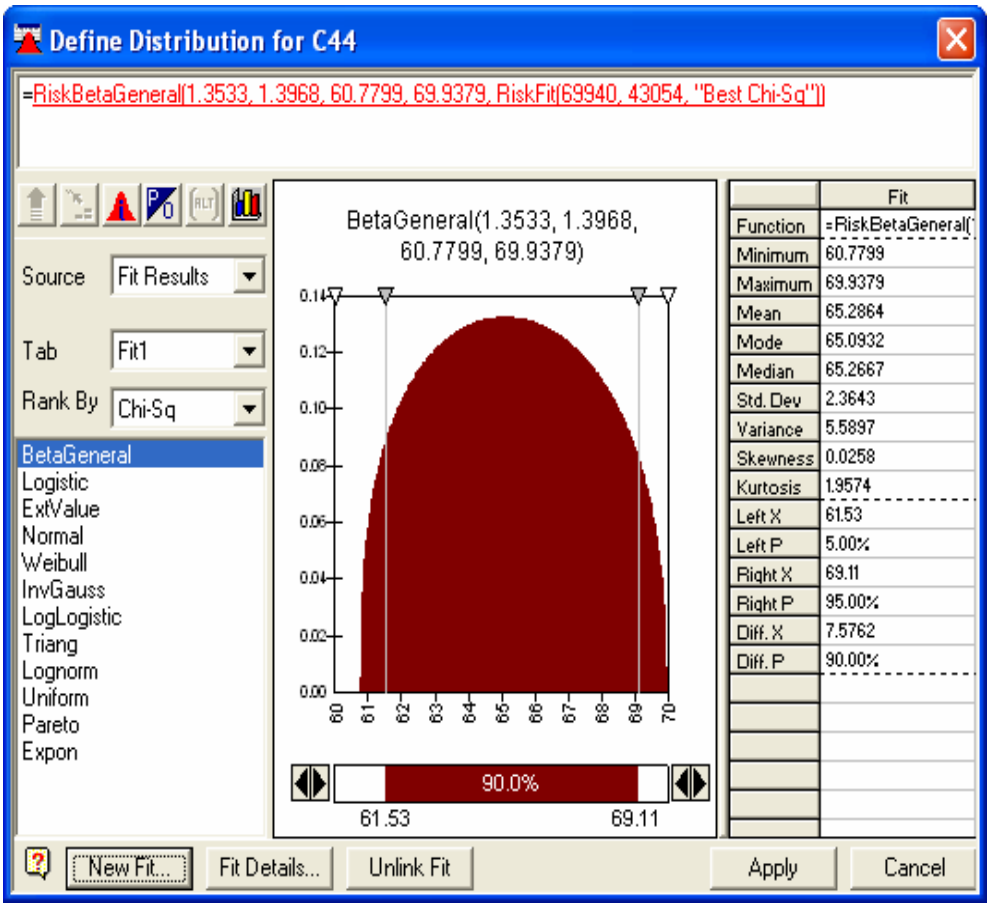


Step 9—The window titled **Define Distribution for C44** will reappear (see Figure 4.5). At the center of this window is a graph with the probability distribution that has been fitted to the data (in this example, Cells C9 through C43) and is ranked the highest based on a Chi-square test. Statistics about the distribution such as mean, variance, 95 percentile, etc. appear at the right of the window. Other possible distributions can be reviewed by clicking on them at the left of the window. They can also be ranked by other tests by clicking on the **Rank By** dropdown box, but for the purposes of this example, we will use the Chi-square test.

At this point, the distribution should also be reviewed for reasonability. For example, if the recommended distribution contains values that are unrealistic for the PMPM figures, the distribution may need to be revised and refitted. Alternatively, @RISK offers a function called **RiskTruncate** that allows you to use the fitted distribution, but exclude values outside a minimum-maximum range when sample values are generated from the fitted distribution. For more on the properties and uses of particular probability distributions, there are a number of excellent references that can be accessed on the Internet.

Click the **Apply** button in the window when you are satisfied with the distribution.

Figure 4.5
The Define Distribution Window after the Range Has Been Fitted



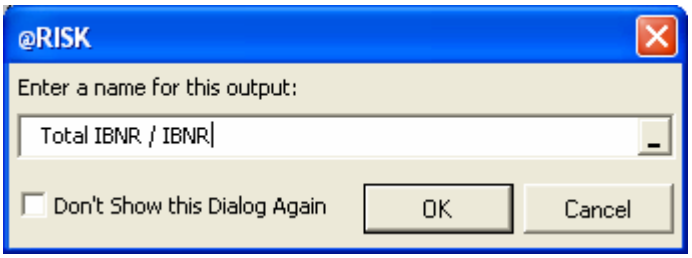
Note: This distribution is ranked highest based on the Chi-square test.

Step 10—Cell C44 now contains the @RISK formula that describes the fitted probability distribution that will be used in the simulation to generate values for that cell.

Step 11—Repeat Steps 1–10 for the remaining columns starting with Cell D43. When the @RISK formula for the fitted probability distribution for Cell D43 has been created, copy that formula down to the remaining cells in the column (Cell D44). Continue the process until all remaining colored columns are completed.

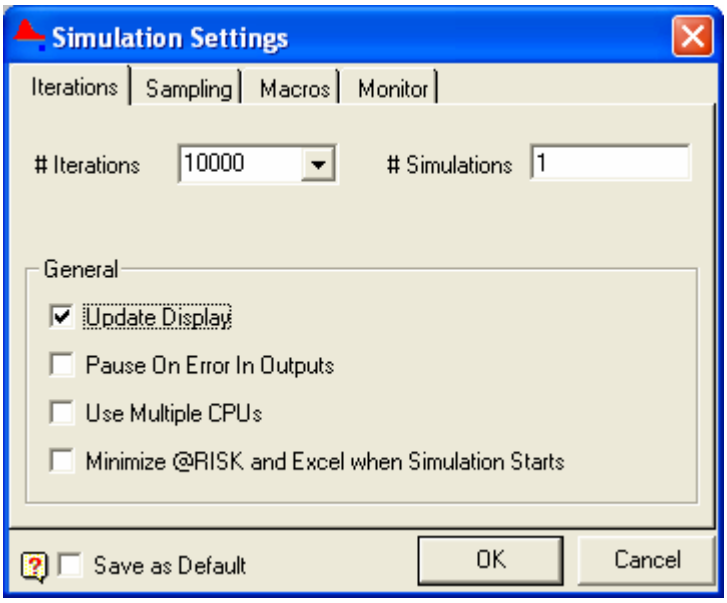
Step 12—Right click on Cell M45. Move to the @RISK selection and choose **Add Output** from the @RISK subwindow that appears (see Figure 4.6). Enter a name for the output (such as Total IBNR) and click on the **OK** button. For the simulation, @RISK will record the value of the output for each iteration. This step has now specified that Total IBNR is the output that we want to record.

Figure 4.6
Add Output Window to Specify Total IBNR



Step 13—On the @RISK tool bar, click on the **Simulation Settings** button. The **Simulation Settings** window appears (see Figure 4.7). Enter the number of iterations by selecting one of the dropdown choices or typing in the number in the **# Iterations** box, say 10,000. Check the **Update Display** box. Click **OK**.

Figure 4.7
Simulation Settings Window



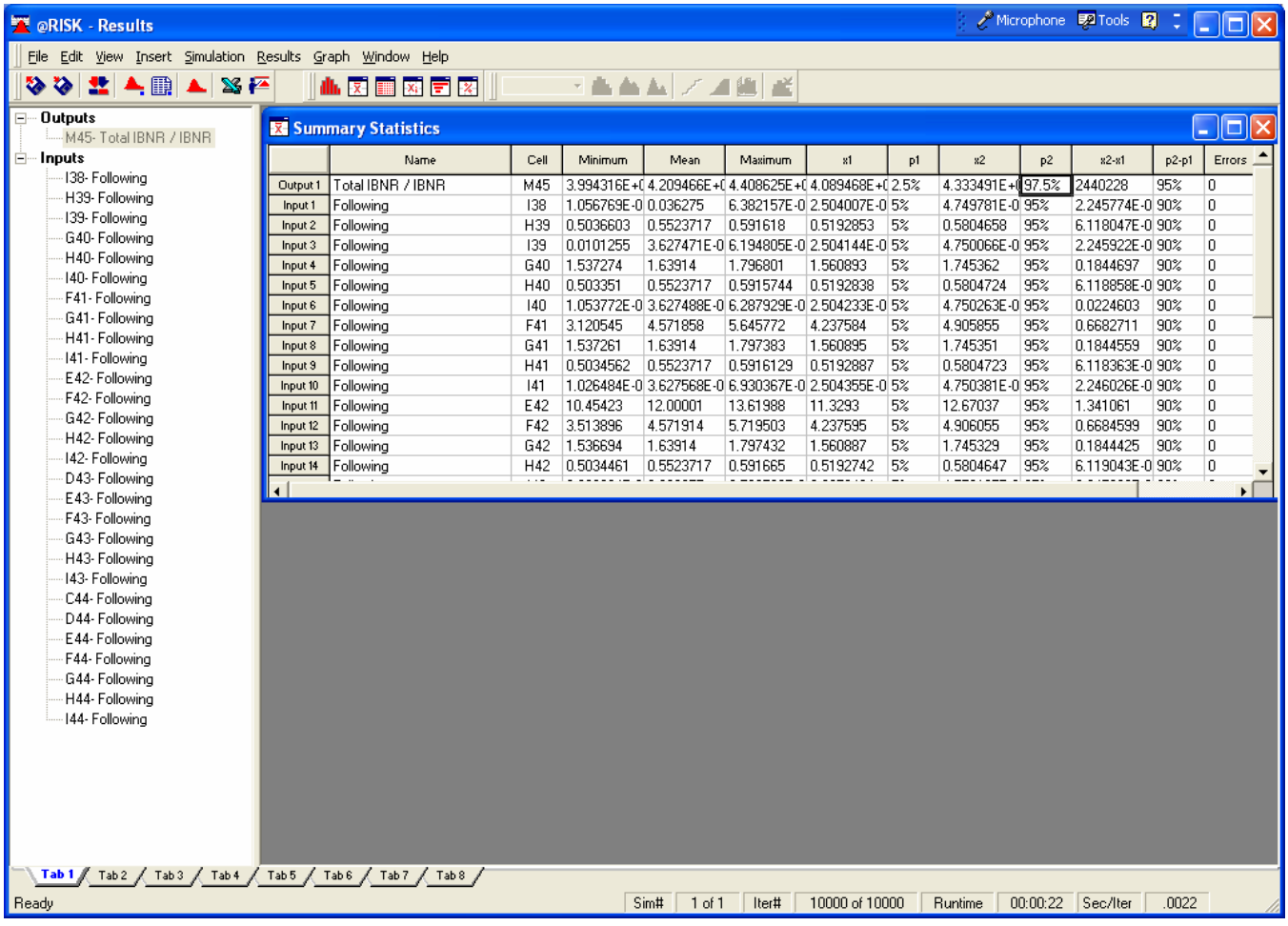
Step 14—Click the **Start Simulation** button on the @RISK toolbar. The simulation starts and the spreadsheet display is updated for each iteration. When all iterations are completed, the @RISK **Summary Statistics** sheet appears (see Figure 4.8). It should be noted that because the **Summary Statistics** sheet displays simulation results, the values seen by the user are unlikely to be exactly the same as in Figure 4.8.

Step 15—The @RISK **Summary Statistics** sheet lists each of the output and input cells used in the simulation. For each cell, the following statistics appear: minimum, mean, maximum and two

user-changeable percentile values and percentile choices, (x1 and x2, p1 and p2, respectively), and the difference between the percentile values (x2-x1, p2-p1).

Step 16—From the information in the **Summary Statistics** sheet, varying levels of confidence intervals around the mean can be calculated. For instance, to calculate a 95 percent interval of the Total IBNR, enter p1 of 2.5 percent and p2 of 97.5 percent into the Total IBNR line that appears in the **Summary Statistics** sheet. The percentile values are automatically updated and the difference between x2 and x1 represents the length of the interval. Other statistics about the Total IBNR can be obtained from the **Detailed Statistics** sheet that can be obtained from the @RISK toolbar.

Figure 4.8
Summary Statistics Window Showing the Output Values for Total IBNR



Note: Columns x1 and x2 from a 95 percent confidence interval for Total IBNR.

4.4 Practical Considerations

Some practical considerations when performing the simulation include:

- Carefully review the range of values that can be generated by each fitted distribution to ensure it is reasonable. For example, if negative claim values or large claim outliers were used to fit a distribution, spurious results may be generated during the simulation. Adjust the data manually or use the @RISK RiskTruncate function to eliminate unreasonable input values.
- If claims have increased rapidly over time, the fitted distributions may underestimate the impact of the increase. The results should be reviewed carefully for such patterns and adjusted accordingly.
- The number of iterations run will impact both the speed of the simulation and the accuracy of the results. For most Total IBNR calculations, the complexity of the calculations can be readily handled in Excel. As such, a high number of iterations (10,000 or more) should be able to be performed rapidly.

4.5 Summary

In this section, we presented an illustrative example of how to apply simulation to the confidence intervals on the total IBNR. Simulation can be applied at differing levels of complexity depending on data availability and the IBNR calculation methods. The greater the complexity of the underlying simulation calculations, the more important it is to run a sufficient number of iterations to ensure robust results.

Section 5—Key Statistical Terms

For reference purposes, this section contains brief explanations of the key statistical terms and concepts presented in the previous sections.

Mean

For a set of data, the mean is the arithmetic average of the data values, equal to the sum of all values divided by the number of values. For a random variable, the mean is its expected value.

For a set of data $x_1, \ldots, x_n$,

$$\text{Mean} = \frac{\text{Sum of Data Values}}{\text{Number of Data Values}} = \frac{\sum x_i}{n} = \bar{x}$$

where $x_i's$ are points of data and n is the number of points of data.

For discrete random variables, the mean equals the sum of each value multiplied by its probability

$$E(X) = \sum x_i \cdot p(x_i) = \mu$$

where n is the number of values X can take (the weighted sum is over all n values) and $p(x_i)$ is the probability that X takes on the value x_i.

For continuous random variables, the mean equals the integral of x (i.e., a value of the random variable) multiplied by the density function. For non-negative random variables, the mean equals the integral from zero to infinity of the survival function of the random variable.

Variance and Standard Deviation

Two sets of data can have the same mean, but be composed of different values. One way to describe this difference quantitatively is to use a measure of dispersion which represents the amount of variation in a data set. This measure is known as the variance for a random variable or the sample variance for a set of data. For a data set, the sample variance is calculated as

$$Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

It is important to note the sample variance is in units that are the square of the original unit. If we were to take the square root of Var(X), we would have the sample standard deviation, s.

For a random variable, its variance is calculated as

$$\text{Var}(X) = E\left(\left(X - E(X)\right)^2\right) = E\left((X-\mu)^2\right)$$

and the standard deviation is $\sigma = \sqrt{\text{Var}(X)}$.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion of a random variable. It relates the mean and standard deviation, and is calculated as the standard deviation divided by the mean:

$$\frac{\sqrt{Var(X)}}{E(X)} = \frac{\sigma}{\mu}.$$

In terms of IBNR estimates, the greater the coefficient of variation, the greater the probability that actual claim runout will be materially different from the best estimate. Conversely, the lower the coefficient of variation, the lower the probability that actual experience, as it emerges, will be materially different from the best estimate.

Confidence Interval

A confidence interval provides an estimated range of values which is likely to include an unknown population parameter. The estimated range is calculated from a given set of sample data. If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage of the intervals will include the unknown population parameter. Confidence intervals are usually calculated so that this percentage is 95 percent, but any percentage can be used. The width of the confidence interval provides some indication of the level of uncertainty about the estimate of the unknown parameter. A very wide interval may indicate that more data should be collected before definitive statements about the parameter can be made.

Prediction Interval

A prediction interval bears the same relationship to a future observation that a confidence interval bears to an unobservable population parameter. Prediction intervals predict the distribution of individual points, whereas confidence intervals estimate the true population mean or other quantity of interest that cannot be observed. For calculating an interval around an IBNR estimate, a prediction interval is normally the measure that is calculated, in technical terms.

Deterministic Method

Deterministic methods produce values for a dependent variable based on independent variables that do not assume any inherent randomness in those variables. Such a method is usually formula or algorithm based, and it does not use statistical techniques such as regression, time series or confidence intervals. For example, the standard application of the completion factor method is a deterministic method.

Statistical Method

Statistical methods account for randomness and uncertainty in observations. In terms of IBNR estimates, a method that assumes claim values follow a probability distribution is considered a statistical method.

Linear Regression

This is a technique in which a straight line is fit to a set of data points to measure the effect of a single independent variable on the value of a dependent variable. The slope of the line is the measure of the impact of the independent variable. The regression technique, in general, seeks to derive the expected value of a random variable Y (dependent variable) given a value of an independent random variable X. The technique is rooted in the observation that if X and Y have the bivariate normal distribution, then the conditional mean of Y given that $X = x$ is a linear function of x. This is a basic methodology for building practical statistical models. Estimation of parameters of the line is done by minimizing the mean square error, i.e., the average of the square of the distance between the value of Y obtained from the actual data and the value of it predicted by the linear relationship. Standard linear regression uses one dependent variable Y and one independent random variable X. Its generalization, called multiple regression, assumes a more complicated linear relationship of the form $Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$.

Modeling Process

The modeling process is a procedure for taking data and applying a logical structure to it in order to replicate certain situations, produce statistical estimates, or predict future conditions. In calculating IBNR, a model is often used to estimate incurred claim amounts by replicating the underlying process and properties of claim payments. In this context, the modeling process is a series of steps for deciding which underlying model is most appropriate for estimating incurred claims. This process is outlined in Section 3.

Exponential Regression

This is a regression technique that produces an exponential curve as the best fit of data composed of values of the independent random variable X and dependent random variable Y. This technique effectively applies linear regression methodology to $\ln Y$ and X. For an exponential form of regression, the following relationships are effectively assumed:

$$Y = Ar^x$$

$$\ln Y = \ln A + X \ln r.$$

thus, the exponential curve is transformed to a linear equation where the slope is $\ln r$ and the y-intercept is $\ln A$ and we can use linear regression to find the slope and the intercept.

Quadratic Regression

This is a regression technique that produces a quadratic relationship of the form $Y = aX^2 + bX + c$ as the best fit of a set of data composed of values of the independent random variable X and dependent random variable Y. This is a generalization of the linear relationship of the form $Y = aX + b$ assumed in linear regression.

Multicolinearity

Multicolinearity is the degree of correlation between independent variables. A high degree of multicolinearity produces unacceptable uncertainty (large variance) in regression coefficient estimates. Specifically, the coefficients can change drastically depending on which terms are included in the model and also the order they are placed in the model. When applied to a model of the form $Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$, multicolinearity of the model means that some of the independent variables $X_1, X_2, \ldots, X_n$ are superfluous, because of their dependence on other variables. In other words, the superfluous variables could be fully predicted by those other variables that they are dependent upon. Multicolinearity is usually produced in multiple regression models by adding too many variables to the model, without consideration for the possibility of those variables being dependent on each other. Thus adding variables to a model would not necessarily produce a better predictive model, or even produce a model that has the same predictive ability as the original model, but in fact might produce a worse model.

Outlier

An outlier is an observation that is far removed from the general pattern of the data. For the purposes of IBNR estimation, the determination of an outlier depends on the model used to estimate the IBNR. An outlier with respect to one model may not be an outlier relative to another model.

Test Statistics and p-Values

For a regression analysis, there are several associated test statistics and corresponding p-values. These are normally included as output from a statistical package. The F-statistic and corresponding p-value test if the overall model is statistically significant. A p-value less than 0.05 is an indicator that the overall model is statistically significant when compared to a model using none of the independent variables. The t-statistics and corresponding p-values test each individual regression coefficient (beta value). The independent (predictor) variables with p-values less than 0.05 are considered to be useful in describing the relationship between the independent (predictor) variables and the dependent (response) variable.

R-Square and Adjusted R-Square Values

When regression models are developed, among the first results reviewed are the R-square and adjusted R-square values.

R-square and adjusted R-square are defined as follows:

R-square is the name of the coefficient of determination of the regression model. It can be interpreted as the percentage of the variation in the observed values of the dependent variable that is explained by the regression model. The larger the R-square value, the greater is the indication that the model is satisfactory. R-square is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

where *SSR* is the sum of squares accounted for in the regression model, *SST* is the total sum of squares of the deviations of the dependent values from their mean, and *SSE* = *SST* - *SSR*. One drawback of R-square is that adding more independent variables causes the R-square value to increase even though there is no significant improvement to the model. In order to correct for this effect, an adjustment can be made to the R-square formula. The adjusted R-square is defined to be

$$R^2(adj) = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)},$$

where *k* is the number of predictors in the regression model and *n* is the total number of observations on the response variable. Larger adjusted R-square values indicate a better fit of the model.

Final Comments

This document has presented approaches for incorporating statistical techniques into actuarial methods used for calculating medical liabilities. These approaches will hopefully motivate practitioners to consider the use of statistical approaches as a way to enhance their current practice. As noted throughout the document, the techniques described are not intended to represent the state of the art in the application of statistics to IBNR calculations. Rather, it is the authors' wish that this guide serve to stimulate further research and development of innovative techniques for more accurate IBNR predictions.

Appendix I—Multicolinearity

In this section, we discuss the issue of multicolinearity, which can occur when the model contains two independent variables that are closely correlated.

I.1 Multiple Regression Model: Correlation between Variables

Let us say there are two linear regression models with the following independent variables:

Model 1: X_1
Model 2: X_1, X_2

In many cases, Model 2 will provide an estimate that is better (or at least as good) as Model 1. However, there are situations where the estimates from Model 2 could be unstable and misleading due to the correlation between X_1 and X_2. This situation occurs in what is known as multicolinearity.

In a multiple linear regression model, when two or more independent variables are highly correlated, multicolinearity or colinearity is present in the model. When multicolinearity occurs, the estimates of the regression coefficients are very unstable, and the predictions made using the model could be misleading.

Multicolinearity is defined to be the existence of near linear relationships between the independent variables. P-values reported in the regression output cannot be used to detect the presence of the colinearity. However in the presence of high colinearity it sometimes occurs that almost all the p-values corresponding to individual coefficients (independent variables) are relatively large indicating that these variables are of no use in predicting the dependent variable, while the p-value corresponding to the F-statistic for the effectiveness of the overall model is small. This normally happens when there is a severe case of multicolinearity, but not in all instances.

Exhibit I.1 provides an illustration of a data set with an extreme case of multicolinearity.

Exhibit I.1
Data Set With Extreme Multicolinearity

Y	X_1	X_2	X_3
77.2163	0.9853	13.6	58.4385
80.4335	1.0938	16.6	65.2210
73.3477	0.9280	10.2	49.7600
68.3209	0.9462	5.4	46.6790
71.4829	0.8885	15.0	54.9825
73.9800	1.0267	9.0	55.0015
76.3007	0.9225	12.3	53.9125
73.9351	0.9372	16.3	58.7740
75.0595	0.8858	15.4	54.7610
78.2235	0.9643	13.0	57.0935
77.8391	0.9316	14.4	57.0220
73.5727	0.9705	10.0	51.4725
82.4024	1.1240	10.2	61.9800
73.1462	0.8517	9.5	47.3265
64.9958	0.7851	1.5	36.3295
80.6591	0.9186	18.5	58.3370
84.3041	1.0395	12.6	59.5775
83.1205	0.9573	17.5	58.8785
67.4872	0.9106	4.9	45.4770
79.7789	1.0070	15.9	59.8150
69.7220	0.9806	8.5	51.4270
77.1627	0.9693	10.6	52.2185
72.1725	0.9496	13.9	57.9320
85.8880	1.1184	14.9	62.3280

This data set contains 24 observations. We will use the variable Y as the dependent (response) variable and X_1, X_2 and X_3 as independent variables.

Exhibit I.2 is a 3-D graph illustrating the correlation of the independent variables.

Exhibit I.2

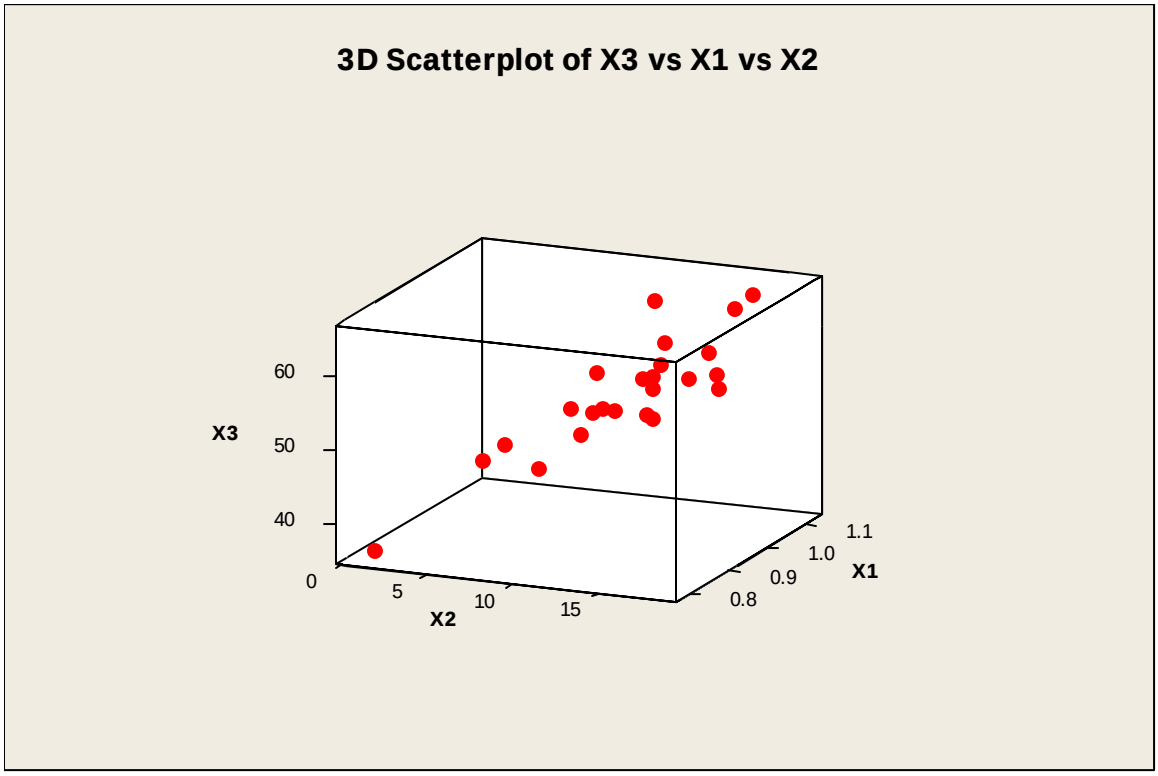


Exhibit I.2 shows that the variables X_1 and X_2 are positively correlated with X_3 and the scatterplots of X_3 versus X_1 (Exhibit I.4) and X_3 versus X_2 (Exhibit I.5) confirm this.

Exhibit I.3 is a scatterplot of X_1 versus X_2.

Exhibit I.3

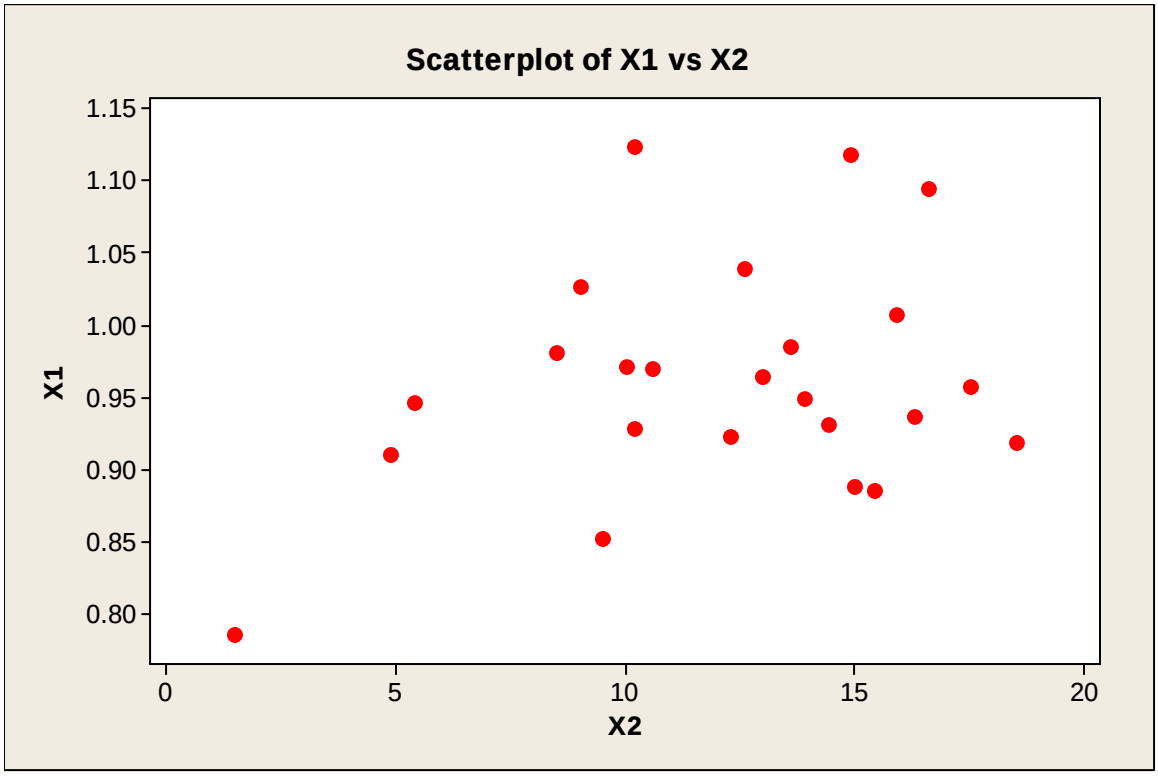


Exhibit I.3 shows that the variables X_1 and X_2 are not highly correlated.

Exhibit I. 4 is a scatterplot of X_1 versus X_3.

Exhibit I.4

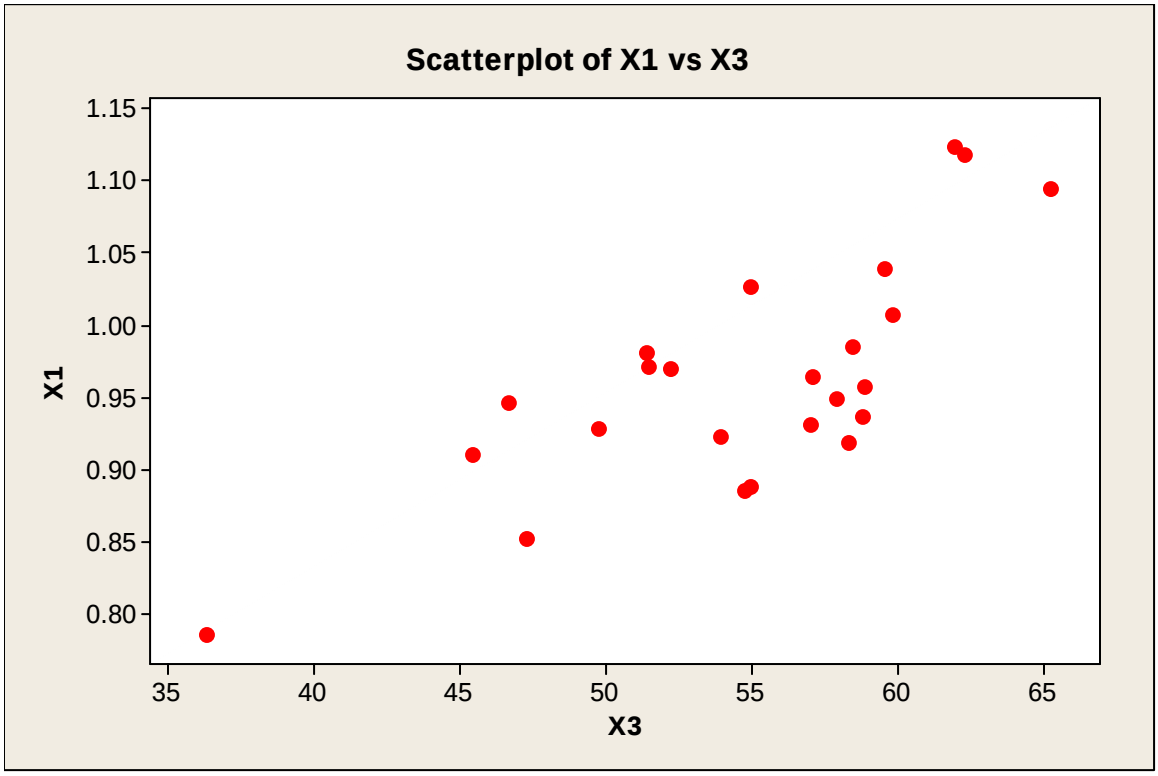


Exhibit I.4 clearly shows that the variables X_1 and X_3 are highly correlated.

Exhibit I.5 is a scatterplot of X_2 versus X_3.

Exhibit I.5

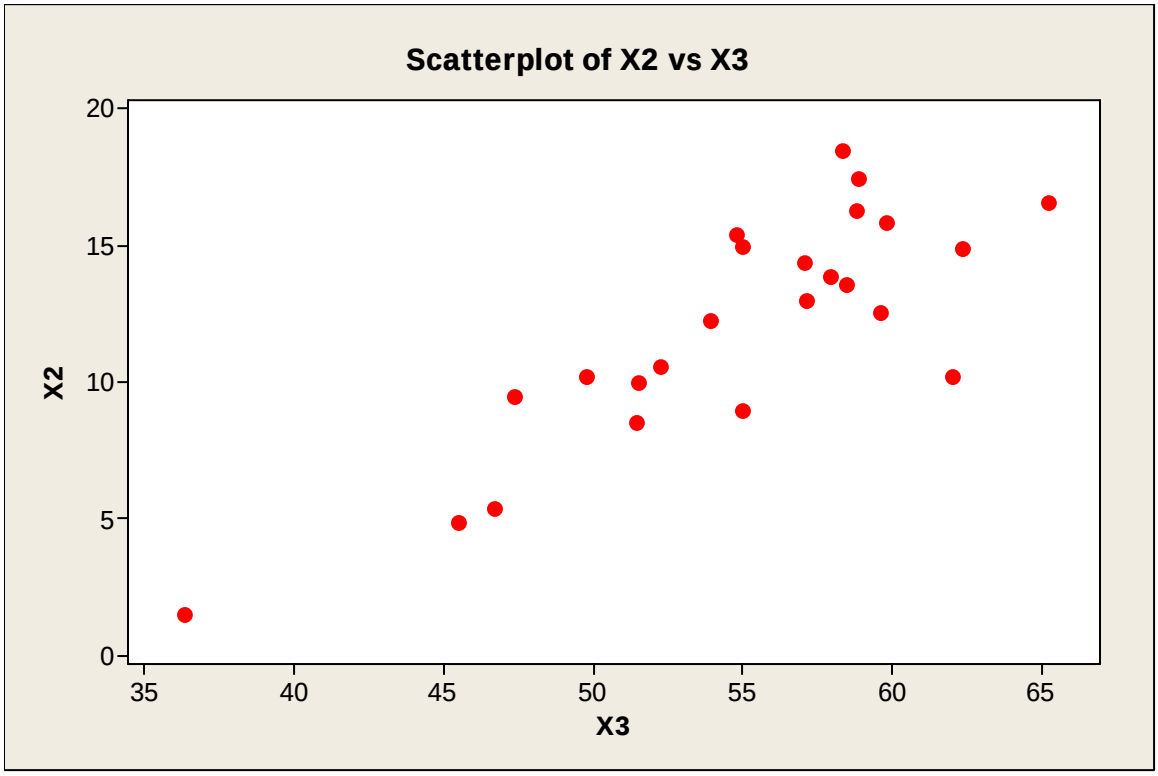


Exhibit I.5 also shows that the variables X_2 and X_3 are highly correlated.

The correlation coefficients between X_1, X_2 and X_3 shown in Exhibit I.6 confirm the scatterplots.

Exhibit I.6
Correlation Coefficients Between Independent Variables

	X_1	X_2
X_2	0.310	---
X_3	0.752	0.837

It is evident X_3 is highly correlated with both X_1 and X_2.

The results of several multiple linear regression models representing differing combinations of independent variables are shown in Exhibit I. 7. Notice that there are several relatively large p (probability) values in some of the models. (A value of less than 0.05 indicates that the independent variable is useful in explaining the values of the dependent variable.)

Exhibit I.7
Comparison of Key Values from Models

	Model				
	1	2	3	4	5
Independent Variables	X_1, X_2	X_1, X_3	X_2, X_3	X_3	X_1, X_2, X_3
R-Square (Adjusted) Value	71.4%	67.2%	66.2%	67.7%	70.6%
P-Values:					
Intercept	0.000	0.000	0.000	0.000	0.000
X_1	0.000	0.419	N/A	N/A	0.056
X_2	0.000	N/A	0.936	N/A	0.081
X_3	N/A	0.001	0.001	0.000	0.526

The detailed regression data can be found in Appendix I-A.

None of the independent variables are significant in Model 5 based on the 0.05 threshold. Only the intercept term is significant. This indicates that the model implies that no independent variable is useful in predicting the values of the dependent variable. Furthermore, this would imply that we might rather use the sample mean of the dependent variable to predict any value of the dependent variable. As can be seen from Model 1, this is completely misleading. This is because the estimators of the coefficients in Model 5 are very unreliable and their errors are very large.

Let us look at the effect of multicolinearity on the prediction. When X_1=0.8 and X_2=2, the predicted value of Y using Model 1 is 63.075, which is close to the minimum value of Y in the data, 64.9958. When X_1=0.8, X_2=2, and X_3=60, the predicted value of Y using Model 5 is 55.178. This value is significantly out of range of the Y-values even though the values of the independent variables are in the range of the observations in the data. This is the consequence of the presence of multicolinearity. In this case just by removing the variable X_3 from the model, we obtain a good model.

In our extreme example, adding a variable made our model poor, and it is evident that adding variables could sometimes make models significantly worse. P-values greater than 0.05 are an indicator that there may be the presence of multicolinearity.

For a more detailed description of multicolinearity, we recommend the text by Belsey, Kuh and Welsch, which is listed in the references section.

Appendix I-A—Regression Models Used in Multicolinearity Analysis

In this section, we present the regression models used in Appendix I to demonstrate multicolinearity.

Our first regression analysis is Y versus X_1 and X_2 (Model 1).

The regression equation is

$$Y = 33.0 + 35.8\ X_1 + 0.692\ X_2.$$

Exhibit I-A.1 illustrates the regression output from Model 1.

Exhibit I-A.1
Regression Output
Model 1

		Coefficient	SE Coef	t	p
Independent Variable	Constant	33.032	7.404	4.46	0.000
	X1	35.823	8.043	4.45	0.000
	X2	0.692	0.151	4.60	0.000

Standard Deviation (S) = 2.913413
R-Square (R-sq) = 73.9%
R-Sq Adjusted (R-sq(adj)) = 71.4%

The analysis of variance for Model 1 is shown in Exhibit I-A.2.

Exhibit I-A.2
Analysis of Variance
Model 1

	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Value (F)	p-Value (p)
Source					
Regression	2	503.866	251.933	29.68	0.000
Residual Error	21	178.247	8.489		
Total	23	682.114			

We can see that all the variables are significant and the adjusted R-square value is 71.4 percent. In addition, the overall F-test for the model is highly significant. These are all indications of a good model fit.

Now we review the results of the multiple linear regression model of Y on all three independent variables X_1, X_2 and X_3.

First, we evaluate the regression analysis: Y versus X_1 and X_3; this is Model 2.

The regression equation is

$$Y = 32.6 + 10.2\ X_1 + 0.610\ X_3.$$

Exhibit I-A.3 illustrates the regression output from Model 2.

Exhibit I-A.3
Regression Output
Model 2

		Coefficient	SE Coef	t	p
Independent variable	Constant	32.600	7.918	4.12	0.000
	X1	10.234	12.408	0.82	0.419
	X3	0.610	0.154	3.97	0.001

Standard Deviation (S) = 3.117639
R-Square (R-sq) = 70.1%
R-Sq Adjusted (R-sq(adj)) = 67.2%

The analysis of variance for Model 2 is shown in Exhibit I-A.4.

Exhibit I-A.4
Analysis of Variance
Model 2

	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Value (F)	p-Value (p)
Source					
Regression	2	478.001	239.000	24.59	0.000
Residual Error	21	204.113	9.720		
Total	23	682.114			

From the exhibits we observe that X_1 has no significant effect in explaining the dependent variable Y in the presence of the independent variable X_3.

Next, we review the regression analysis of Y versus X_2 and X_3; this is Model 3.

The regression equation is

$$Y = 37.7 + 0.023\ X_2 + 0.692\ X_3.$$

Exhibit I-A.5 illustrates the regression output from Model 3.

Exhibit I-A.5
Regression Output
Model 3

		Coefficient	SE Coef	t	p
Independent variable	Constant	37.656	7.685	4.90	0.000
	X_2	0.023	0.285	0.08	0.936
	X_3	0.692	0.188	3.68	0.001

Standard Deviation (S) = 3.167226
R-Square (R-sq) = 69.1%
R-Sq Adjusted (R-sq(adj)) = 66.2%

The analysis of variance for Model 3 is shown in Exhibit I-A.6.

Exhibit I-A.6
Analysis of Variance
Model 3

	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Value (F)	p-Value (p)
Source					
Regression	2	471.456	235.728	23.50	0.000
Residual Error	21	210.658	10.031		
Total	23	682.114			

From the exhibits we observe that the variable X_2 has no significant effect in explaining the dependent variable Y in the presence of the independent variable X_3.

The next model, Model 4, is a regression analysis of Y versus X_3.

The regression equation is

$$Y = 37.2 + 0.705\ X_3.$$

Exhibit I-A.7 illustrates the regression output from Model 4.

Exhibit I-A.7
Regression Output
Model 4

		Coefficient	SE Coef	T	p
Independent variable	Constant	37.232	5.542	6.72	0.000
	X_3	0.705	0.101	7.02	0.000

Standard Deviation (S) = 3.094900
R-Square (R-sq) = 69.1%
R-Sq Adjusted (R-sq(adj)) = 67.7%

The analysis of variance for Model 4 is shown in Exhibit I-A.8.

Exhibit I-A.8
Analysis of Variance
Model 4

	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Value (F)	p-Value (p)
Source					
Regression	1	471.389	471.389	49.21	0.000
Residual Error	22	210.725	9.578		
Total	23	682.114			

Since the adjusted R-square for Model 4 is 67.7 percent versus 71.4 percent for Model 1, Model 1 is better in explaining the dependent variable Y.

As a final illustration of the effects of multicolinearity, we will review the Model 5 results. Model 5 is a regression of Y versus X_1, X_2 and X_3.

The regression equation is

$$Y = 33.1 + 51.2\ X_1 + 1.05\ X_2 - 0.350\ X_3.$$

Exhibit I-A.9 illustrates the regression output from Model 5.

Exhibit I-A.9
Regression Output
Model 5

		Coefficient	SE Coef	t	p
Independent variable	Constant	33.113	7.510	4.41	0.000
	X1	51.234	25.213	2.03	0.056
	X2	1.047	0.569	1.84	0.081
	X3	-0.350	0.542	-0.65	0.526

Standard Deviation (S) = 2.954676
R-Square (R-sq) = 74.4%
R-Sq Adjusted (R-sq(adj)) = 70.6%

The analysis of variance for Model 5 is shown in Exhibit I-A.10.

Exhibit I-A.10
Analysis of Variance
Model 5

	Degrees of Freedom	Sum of Squares	Mean Square	F-Value	p-Value
Source	(DF)	(SS)	(MS)	(F)	(p)
Regression	3	507.512	169.171	19.38	0.000
Residual Error	20	174.602	8.730		
Total	23	682.114			

Notice that the overall F-test is highly significant and the adjusted R-square value is 70.6 percent, which is comparable to the value obtained for Model 1, but Model 5 is not an appropriate model for our purposes.

References

Key references in the study of statistical methods for computing medical insurance IBNR include the following:

Barnett, G., and B. Zehnwirth. 2000. Best estimates for reserves. *Proceedings of the Casualty Actuarial Society* 87, pt. 2, no. 167, Nov. 12–15, pp. 245–321. Available online at: http://www.casact.org/pubs/proceed/proceed00/00245.pdf.

Belsey, D.A., E. Kuh, and R.E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Black, K., Jr., and H.D. Skipper, Jr. 2000. *Life and Health Insurance,* 13th ed. Upper Saddle River, N.J.: Prentice Hall.

Bornhuetter, R.L., and R.E. Ferguson. 1972. The actuary and IBNR. *Proceedings of the Casualty Actuarial Society* 59, no. 112, pp. 181–195.

Brown, R.L. 2001. *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance*, 2nd ed. Winsted, Conn.: Actex Publications.

Fearrington, D., and R. Lynch. 2004. Approaches to determining unpaid claim liabilities: Old and new. *The Record of the Society of Actuaries,* Valuation Actuary Symposium, Boston, Mass., Sept. 20–21, Session 39TS.

Litow, M.E. 1989. A modified development method for deriving health claim reserves. *Transactions of the Society of Actuaries* 41.

Little, R.J.A., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley.

Lloyd, J.C. 2005. Health reserves. *Society of Actuaries Study Note 8GM-305-00*. Schaumburg, Ill.: Society of Actuaries.

Lynch, R.G. 2004. Gerbils on espresso: A better way to calculate IBNR reserves with low variance. *Contingencies*, Jan.-Feb., pp. 28–38.

Mosley, R.C. 2004. Estimating claim settlement values using GLM. *Casualty Actuarial Society, Discussion Paper Program—Applying and Evaluating Generalized Linear Models Including Research Papers on the Valuation of P&C Insurance Companies.* Available online at: http://www.casact.org/pubs/dpp/dpp04/.

Pindyck, R.S., and Rubinfeld, D.L. 1998. *Econometric Models and Economic Forecasts*, 4th ed. New York: Irwin McGraw-Hill.

Rubin, D.B. 1986. *Multiple Imputations for Non-Response in Surveys*. New York: John Wiley.

Sen, A., and Srivastava. M. 1990. *Regression Analysis: Theory, Methods, and Applications*. New York: Springer-Verlag.

Society of Actuaries. 2006. Course 7, March, Advance Reading.

Sutton, H.L., and Sorbo, A.J. 1987. *Actuarial Issues in the Fee-For-Service/Prepaid Medical Group*. Denver, Colo.: Center for Research in Ambulatory Health Care Administration, pp. 62–67.

Zehnwirth, B. 1994. Probabilistic development factor models with applications to loss reserve variability, prediction intervals, and risk based capital. *Casualty Actuarial Society Forum*, pp. 510–605.
Available online at: http://www.casact.org/pubs/forum/94spforum/94spf447.pdf.

Zehnwirth, B. 1996. Three powerful diagnostic models for loss reserving. University of Melbourne, Centre for Actuarial Studies, Australia, Working Paper Series: http://www.economics.unimelb.edu.au/actwww/No34.pdf.